AD-A222 205

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1989	Final Scientific 1 Mar 86 - 30 Nov 89

4. TITLE AND SUBTITLE
NUMERICAL SOLUTION TO STEADY-STATE PROBLEMS WITH DISCONTINUITIES

5. FUNDING NUMBERS
AFOSR-86-0127
61102F 2304/A3

6. AUTHOR(S)
David Sidilkover, under the supervision of Professor Achi Brandt

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Weizmann Institute of Science
Department of Applied Mathematics
76100 Rehovot
Israel
AFOSR-TR-

8. PERFORMING ORGANIZATION REPORT NUMBER
90-0579

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Sponsoring: Air Force Office of Scientific Research
Bolling AFB, DC 20332-6448
Monitoring: European Office of Aerospace, Research and Development, Box 14, FPO, NY 09510-0200
NM

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFOSR-86-0127

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release;
Distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
The long term systematic advance of multilevel computational methods has continued in 1989. Highly parallelizable multilevel techniques have been developed for the following mathematical tasks: (1) Solution of general nonlinear PDE systems, including elliptic and nonelliptic, steady-state and time-dependent problems and inverse problems, pointing toward new type of multigrid solvers in fluid dynamics. (2) Fast integral transforms, including FFT on non-uniform grids. (3) Solution of integro-differential equations. (4) Fast multiplication by a dense matrix or its inverse, including O(n) calculation and solution of n body interactions. (5) Global optimization of systems with many local optima, including in particular discrete optimization. (6) Linear programming, at least for spatial problems. (7) Image restoration. (8) Computing behavior of statistical fields; fast calculation of thermodynamic limits. (9) Derivation of macroscopic material equations from microscopic physics. (10) Fast calculation and super-fast updating of large determinants of grid equations, with special emphasis on Dirac equations.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Multigrid, numerical PDE, fluid dynamics, time stability, local mode analysis, Integral transform, many body interactions, integral equations, material research, global optimization, image restoration, statistical physics, parallel processing	80
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Numerical Solution to Steady-State Problems with Discontinuities

Thesis for the Degree of

Doctor of Philosophy

by

David Sidilkover

Under the supervision of Prof. Achi Brandt [1]

Department of Applied Mathematics
The Weizmann Institute of Science

Submitted to the Scientific Council of
The Weizmann Institute of Science
Rehovot 76100, Israel

August 1989

[1]Research supported mainly by the Air Force Office of Scientific Research, United States Air Force, under grant AFOSR-86-0127, and also under grant AFOSR-86-0126 and by the National Science Foundation under grant NSF DMS-8704169

This work has been carried out in the Department of Applied Mathematics at the Weizmann Institute of Science, Rehovot, Israel, under the supervision of Professor Achi Brandt.

Acknowledgements

I would like to thank Professor Achi Brandt for guidance, help and patience during the various stages of this work.

I wish also to thank Professor Uri Ascher (UBC, Canada) for reading the manuscript and making remarks, which led to its improvement.

I am also grateful to Dr. S. Spekreijse (currently at NLR, Holland) for some very stimulating discussions.

I wish to thank my colleagues and friends Joe Greenwald, Vladimir Mikulinsky and Irad Yavneh for many helpful discussions.

Моим родителям посвящается.

Abstract

The primary topic of this thesis is the multigrid solution of the steady state 2D nonlinear conservation law. The difficulties encountered in the numerical solution of more complicated problems in fluid dynamics (e.g. the steady 2D Euler system), like the non-ellipticity of the equations, and the presence of discontinuities in the solution, can be studied in this model scalar case.

The work deals with two main issues: development of new discretization schemes and adaptation of the coarsening techniques.

New genuinely 2D (based on a "9-point square" stencil) conservative discretization schemes are developed. These schemes provide the possiblity to separate treatment of the streamwise and cross-stream directions. Due to this separation, the artificial viscosity can be added in the streamwise direction only. High resolution is introduced in the cross-stream direction. Therefore, the resulting schemes have good stability properties, are second order accurate and provide a good resolution of discontinuities: representing them in the numerical solution by thin oscillation-free transition layers.

The adaptation of the coarsening techniques is based upon a more precise understanding of what should be meant by discontinuity location in a shock-captured solution. We have shown that a such solution (provided the discretization scheme employed is conservative and second order accurate) contains information about the discontinuity location with second order accuracy. The conventional coarsening techniques (full-weighted residual transfer and bilinear correction interpolation) appear to provide a good correction for the discontinuity location as well as for the solution in smooth regions.

As the result an efficient multigrid solver is constructed for a general steady state 2D conservation law.

Contents

Chapter 1

Introduction

A great need for accurate simulations of flows with discontinuities exists in many fields of physics and engineering. The major difficulties in this area are the non-ellipticity of the governing equations and the problem of representing the shocks and contact discontinuities.

Multigrid methods were initially developed for elliptic problems and have been shown to be extremely efficient in this case. However, the non-elliptic case still requires investigation in order to achieve the same efficiency.

The objective of this work is to develop a fast multigrid solver for the scalar steady state 2D conservation law. Such a conservation law can serve as a good model problem for more general non-elliptic problems (like the steady-state 2D Euler system). This is because some difficulties encountered in their numerical solution, such as the conflict between numerical stability and higher-order accuracy or the presence of discontinuities in the solution, can be studied already in the scalar case. Therefore, this work can be regarded as a preparational step towards developing a fast solver for more complicated non-elliptic problems.

This solver is required to approximate the solution such a conservation law fast, approaching steady state directly without an explicit marching in time. We want it not only to produce second order accurate solution in smooth regions, but also to deal with discontinuities in some efficient way. Therefore, we explore the problem of representing discontinuities and getting them to converge to their correct position by a multigrid solver as efficiently as obtaining the solution in smooth regions is explored.

One approach to representing discontinuities by a finite difference method is to use difference equations only in smooth regions. The discontinuity itself is followed explicitly using some jump conditions supplemented by characteristic data. This approach is called shock fitting. A multigrid fast solver can be used both for obtaining the solution in smooth regions and for following the path of a discontinuity in an efficient way. It is also possible to show that, once we obtain higher order accuracy in the smooth regions, we can obtain the same order of accuracy in the discontinuity location. However, for complex flows with several intersecting shocks such procedures become difficult. Furthermore, there is the additional difficulty of predicting the generation of shocks that are not present initially.

Another alternative is the so-called shock capturing method. In the flow of a real fluid there are no discontinuities. There are instead very thin regions of very steep gradients.

This is because terms representing viscosity and heat conduction in the usual equations of fluid dynamics have very small but still non-zero coefficients. However, the distance scale on which the resulting transition layers are smooth in the physical solution may be smaller than any reasonable meshsize. The numerical method then is to increase the size of these dissipative terms so that the flow is not distorted much in smooth regions, but the discontinuities are spread over distance scales resolvable on a practical computational mesh.

This approach leads to some difficulties. The width of the transition layer representing a discontinuity in case of a shock is only few meshsizes. However, it grows indefinitely in space with the distance from the initial boundary in case of a contact discontinuity. Another defect is that when the difference scheme is more than first order accurate, this transition is not monotonic anymore. Large numerical overshoots and undershoots occur from both sides of the discontinuity. Whether one needs to remove these oscillations except for aesthetic reasons seems to be problem dependent. For instance, one unpleasant consequence of these oscillations may be that some quantities may attain values out of their physical domain (like negative density or pressure etc.). Another consequence may be that in case of a strong shock (when the characteristic field is essentially convergent) these oscillations may cause convergence to a non-physical solution or even the loss of stability.

There exists a whole family of one-dimensional methods which produce a higher order solution in the smooth regions and give a sharp resolution of discontinuities. They are so-called high resolution schemes, which are shown to be total variation diminishing - TVD (see [11, 12, 16, 21]), and lately essentially non-oscillatory schemes - ENO (see [13]). These methods are successfully used to solve 1D time-dependent problems. However, even TVD schemes are shown to be convergent for finite time only.

The situation with two-dimensional methods is much less favorable. The explanation is probably that the physics of one-dimensional flows is especially simple and well understood, and easy to imitate by a numerical process. Also, it has been shown in [9], that a higher order accurate 2D scheme cannot be TVD. We can conclude that the theoretical results for the TVD schemes may not be relevant for the higher order methods applied to 2D steady state problems, which is our interest here.

One approach toward the numerical simulation of multidimensional flows is called "dimensional splitting". The multidimensional differential operator is expressed as a sequence of one-dimensional operators, which are then approximated by one-dimensional difference operators (a review can be found in [23]). This approach completely ignores the multidimensional nature of the flow (vorticity, the possible use of diagonal neighbors in the difference scheme, the infinite number of wave propagation directions instead of only two in 1D). Therefore it is important to develop fully multidimensional methods.

The key here is probably to imitate the anisotropic nature of the multidimensional fluid flow. This is possible to achieve by using moving grids with gridlines aligning with the stream direction. However, this simple idea turns out to be very difficult to implement even in case of a simple problem. Our approach here is to construct genuinely

2D difference schemes which use a fixed grid, but provide a separation between the treatment of streamwise and cross-stream directions. This allows us to combine in one difference scheme properties of stability with good resolution of discontinuities and higher order accuracy.

We show in this work that, once a discontinuity is recognised in the shock capturing solution, its location can be recovered up to the same order of accuracy as achieved by the numerical solution in smooth regions. This means that using the shock capturing approach one can obtain as much information about a solution as using the shock fitting approach. The multigrid solver apears to be as efficient in converging the discontinuity location as in converging the solution in smooth regions.

The next chapter presents a discussion about the boundary-value problem for the 2D non-linear steady-state conservation law, basic properties of its solution and some considerations regarding its discretization. We show that if conservative second order accurate stable scheme is used, it is possible to recover the discontinuity location from the obtained numerical solution with second order accuracy.

Chap.3 contains a description a traditional multigrid algorithm for elliptic problems as well as the discussion about adaptation of coarsening procedure for the non-elliptic case with disconinuous solutions, assuming that a stable second order accurate discretization is employed by the algorithm.

Chap.4 is devoted to the construction of such a monotonic discretization for the case of a linear constant coefficient equation. The generalization of this discretization for the general non-linear case is presented in Chap.5.

Chap.6 reports about the numerical experiments, where we examine the accuracy of the obtained solutions (both in terms of the solution error in smooth regions and the error in discontinuity location as well as the perforemance of the algorithm.

Chap.7 contains the summary, the efficiency comparison with the existing methods together with the discussion of future possible developments.

Chapter 2

Conservation law and its discretization

A simple differential equation, but typical to more complicated systems in fluid dynamics, is the 2D nonlinear steady-state conservation law

$$-\varepsilon\Delta u + (f(u))_x + (g(u))_y = s(x,y), \tag{2.1}$$

where $\varepsilon > 0$ is small, Δ denotes the laplacian, f, g and s are given functions.

2.1 Boundary-value problem: solution properties

We shall discuss properties of the solution of the boundary-value problem for (2.1) in order to get insight to its discretization.

2.1.1 Linear convection diffusion equation

Consider first a linear version of (2.1) – a convection-diffusion equation

$$-\varepsilon\Delta u + au_x + bu_y = s(x,y), \tag{2.2}$$

The line whose tangent at every point is determined by the vector $(a(x,y), b(x,y))$ is called a characteristic line. The entire domain can be covered by a family of characteristic lines (characteristics). We want to single out the time-like direction along these lines. In order to be consistent with the sign of the elliptic term coefficient the direction of the vector (a,b) must be the choice. We shall call it the characteristic direction or the stream direction.

Consider the degenerate case ($\varepsilon = 0$) of Eq.(2.2). This equation will be hyperbolic with respect to the characteristic direction (a,b). The boundary value problem for this equation is not well posed. On the other hand, suppose we select the part of the boundary, at every point of which the vector $(a(x,y), b(x,y))$ is directed into the domain. Letting the data on this part of the boundary serve as initial data for the hyperbolic equation, the obtained initial-value problem for the hyperbolic equation is well posed, and its solution has the property that the information from the initial data propagates along characteristics. This means that the value of the solution at every particular point of each characteristic line depends only on the initial value at that point of the boundary

where this characteristic line comes from and on the values of the right-hand-side s along this characteristic. Hence, if the initial data are oscillatory, these oscillations will be convected into the domain along characteristics without any damping. We shall call the solution component which is smooth in the streamwise direction, but oscillates otherwise, the characteristic component. If the initial data are discontinuous at some point, the solution of the equation will have a discontinuity (called a contact discontinuity) along the characteristic line which emanates from this point of the boundary.

Return now to the original elliptic (viscous) problem (2.2). There exists a layer of width $O(\frac{\varepsilon}{|a|+|b|})$ (called a boundary layer), along that part of the boundary not taken as the initial data for the hyperbolic problem, where the solution of the elliptic problem may have rapid changes. The solution of the original elliptic problem differs from the solution of the above described hyperbolic problem only by $O(\varepsilon)$ in the whole domain, except for the boundary layer. If the solution of the hyperbolic problem contains a contact discontinuity, the solution of the original elliptic problem will have a fast change smeared over a layer of width $O(r^{\frac{1}{2}}(\frac{\varepsilon}{|a|+|b|})^{\frac{1}{2}})$, where r is the distance along characteristic from the initial data.

The characteristic components of the solution will be dampened as they are convected along characteristics. We want to obtain a quantitative description of this damping. We shall recall the result for the discrete approximation of (2.2) presented in Sec.2 of [4]. The characteristic component with cross-stream wavelength η will lose a substantial fraction of its amplitude when reaching a certain distance $r_{h,p}(\eta)$ from the boundary into the domain. This $r_{h,p}(\eta)$ is called a survival depth of the η component and as it is shown in Sec.2 of [4]

$$r_{h,p}(\eta) = O(\alpha_1 \eta^{p+1} h^{-p}), \tag{2.3}$$

where h is the meshsize, p is the order of approximation, and αh may reflect the "width" of the stencil, i.e. its diameter when projected on a plane perpendicular to the characteristic direction.

2.1.2 Nonlinear equation

Consider now the degenerate case of Eq.(2.1). We shall obtain the hyperbolic equation

$$(f(u))_x + (g(u))_y = s(x,y). \tag{2.4}$$

Rewrite the equation (2.4) in the quasilinear form

$$a(u)u_x + b(u)u_y = s(x,y), \tag{2.5}$$

where

$$\begin{aligned} a(u) &= df/du \\ b(u) &= dg/du. \end{aligned} \tag{2.6}$$

As in the linear case, characteristics may be introduced and the characteristic direction can be determined. The part of the boundary from which the vector (a,b) is directed

into the domain can be selected and the boundary conditions there can be considered as the initial conditions for Eq.(2.4).

Integrate equation (2.4) over domain Ω and apply Gauss theorem

$$\iint_\Omega s\,dx\,dy = \iint_\Omega [(f(u))_x + (g(u))_y]\,dx\,dy = \int_{\partial\Omega} (f,g)_n\,d(\partial\Omega). \qquad (2.7)$$

Here $\partial\Omega$ denotes the boundary of the domain Ω, and $(\,,\,)_n$ the outward normal (to $\partial\Omega$) of the vector $(\,,\,)$. Relation (2.7) expresses conservation: the integral of the flux through the boundary is equal to the integral of the sources inside the domain. u is called a generalized solution of Eq. (2.4) if it satisfies its integral form (2.7), i.e. if (2.7) holds for every smoothly bounded domain Ω.

The solution of the initial value problem for (2.4) as well as for (2.7) may be non-unique and discontinuous even in case of continuous initial data (see [10, 22]). However, we are interested only in the solution which can be obtained from the solution of the boundary value problem for Eq.(2.1) at the limit $\varepsilon \to 0$. Such a limit solution may be discontinuous, but is unique (see [22]).

Consider again Eq.(2.5). The main difference from the linear case is that the position of characteristics depends also on the solution. Therefore, characteristics may intersect, i.e. information from different points of the initial data will be brought along characteristics to the same point, their intersection point. In such a case discontinuity of the solution, which is called a shock wave, will be produced. Only such discontinuities are physically relevant. In other words, through every point of the path of a discontinuity in the (x,y) plane one can draw two characteristics, one on each side of the shock. The discontinuity will be physically relevant if these two characteristics can be traced in the upstream direction back to the initial data. (Contact discontinuity is a limit case of such a situation.)

Introduce the following symmetric vector-valued function of two variables

$$S(u,v) = [\,(f(u),g(u)) - (f(v),g(v))\,]\,\mathrm{sign}(u-v). \qquad (2.8)$$

Let $\Gamma(u) \in \Omega$ be a set of points where the function $u(x,y)$ is discontinuous. Then this discontinuity will be an admissible one (physically relevant) if the following inequality (the entropy condition) is satisfied

$$(S(u^+,c),\nu) \le (S(u^-,c),\nu), \qquad (2.9)$$

where c is an arbitrary constant, ν is the normal to $\Gamma(u)$, u^+ is the limit value of the solution from the side of discontinuity where ν is directed to, u^- the limit value of the solution from the opposite side of the discontinuity ([22]).

The generalized solution of (2.4), which satisfies (2.9) is unique and it coincides with the limit solution (when $\varepsilon \to 0$) of the boundary value problem for (2.1) ([22]).

2.2 Discretization of conservation law

Our approach to the discretization of (2.4) will be guided by its integral form (2.7), because it holds also accross discontinuities.

Consider a computational grid with a meshsize h in both x and y directions, which covers domain Ω. Integrate Eq.(2.4) over computational cell $C_{i,j}$ surrounding gridpoint (i,j) (see Fig. 2.1)

$$\int_{\partial C_{i,j}} (f,g)_n \, d(\partial C_{i,j}) = \iint_{C_{i,j}} s \, dx \, dy. \tag{2.10}$$

The integral in the right-hand-side of (2.10) can be splitted into four parts – one along each segment of the computational cell boundary. Since we are interested in second order accurate solutions, it will be sufficiently accurate to approximate these integrals using the mid-point rule. The 2D mid-point rule can be used to approximate the left-hand-side of (2.10) as well.

Then the discrete form of (2.10) (we shall call it the balance equation) can be written as follows:

$$h\,(f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j} + g_{i,j+\frac{1}{2}} - g_{i,j-\frac{1}{2}}) = h^2\, s_{i,j} \tag{2.11}$$

The difference scheme constructed this way is called conservative in that it has a property analogous to (2.7): in the sum of discrete equations over all the grid points only numerical fluxes through the boundary remain and all other fluxes cancel each other. The method of calculating the numerical fluxes in (2.11) will be presented in Chapters 4,5.

2.3 Recovering of discontinuity location

As well known, the basic advantage of conservative schemes is that when the meshsize $h \to 0$ the discontinuities will converge to their correct location. However, the question of the order of convergence has remained open. Moreover, it was not clear what should be understood by a discontinuity location in case of a discrete solution. We define here a discontinuity location for a shock capturing solution and we show that h^2 convergence in it can be achieved, if there is h^2 convergence in the smooth regions.

2.3.1 Integral relation

Suppose the exact physically relevant solution of (2.4) contains a discontinuity with a path d (see Fig. 2.2).

Suppose we also have a numerical solution of (2.4) obtained by means of a certain conservative finite difference scheme (2.11). The discontinuity will be represented in the numerical solution by a transition layer (monotonous or containing oscillations). Assume that the numerical fluxes approximate the exact fluxes at the corresponding points with accuracy $O(h^2)$ in the smooth regions, away from the influence of the transition layer.

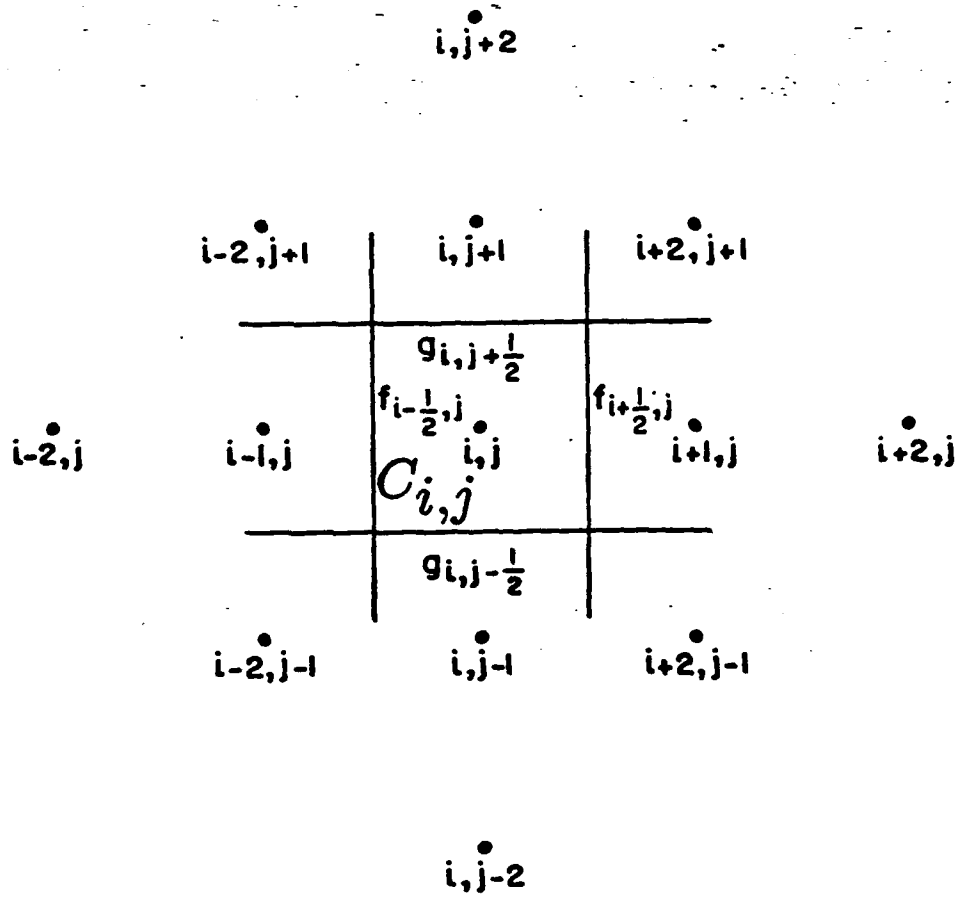

Figure 2.1: Computational grid and computational cell.

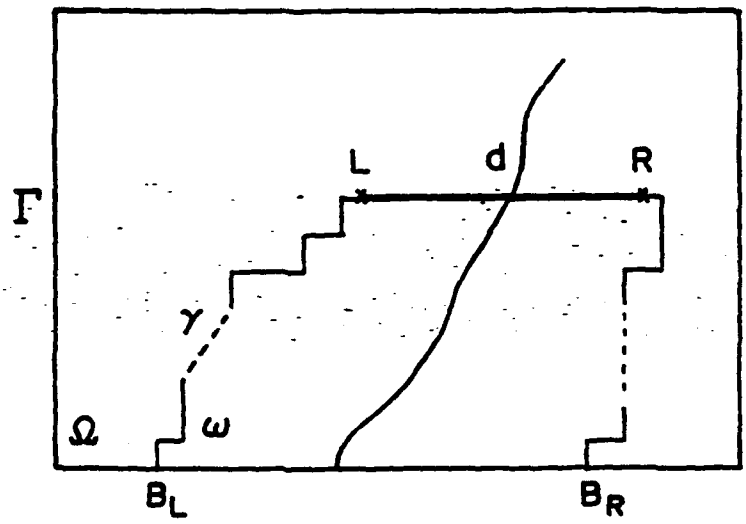

Figure 2.2: Domain and discontinuity path.

Draw a segment LR across the discontinuity. Suppose that this segment goes along computational cell boundaries and that there is already a pointwise h^2 convergence to the exact solution at the point L and R. We want to connect each of these two points with the boundary by a path which will go along computational cell boundary and will belong to the region with h^2 convergence. Keeping these lines close to characteristics going from L and R back to the boundary (Fig. 2.2) seems to be a good choice, because of the physical relevance of the solution. The integration of (2.4) over the subdomain bounded by $B_L L R B_R$ using Gauss theorem will give

$$\int_{\partial\omega} (f,g)_n \, d(\partial\omega) = \iint_{\omega} s \, dx \, dy. \tag{2.12}$$

Split the boundary integral into four parts:

$$\begin{aligned} \int_{\partial\omega} (f,g)_n \, d(\partial\omega) \;&=\; \\ \int_{B_R B_L} (f,g)_n \, d(\partial\omega) \;&+\; \int_{B_L L} (f,g)_n \, d(\partial\omega) + \\ \int_{R B_R} (f,g)_n \, d(\partial\omega) \;&+\; \int_{LR} (f,g)_n \, d(\partial\omega). \end{aligned} \tag{2.13}$$

By (2.11) the summation of the balance equations over the computational cells included in ω is

$$h \sum_{\omega} (f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j} + g_{i,j+\frac{1}{2}} - g_{i,j-\frac{1}{2}}) = h^2 \sum_{\omega} s_{i,j}. \tag{2.14}$$

Notice that in the last summation all numerical fluxes cancel each other except for those through the boundary γ of subdomain ω. Therefore, we denote

$$h \sum_{\partial\omega} (f^h, g^h)_n = h \sum_{\omega} (f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j} + g_{i,j+\frac{1}{2}} - g_{i,j-\frac{1}{2}}). \tag{2.15}$$

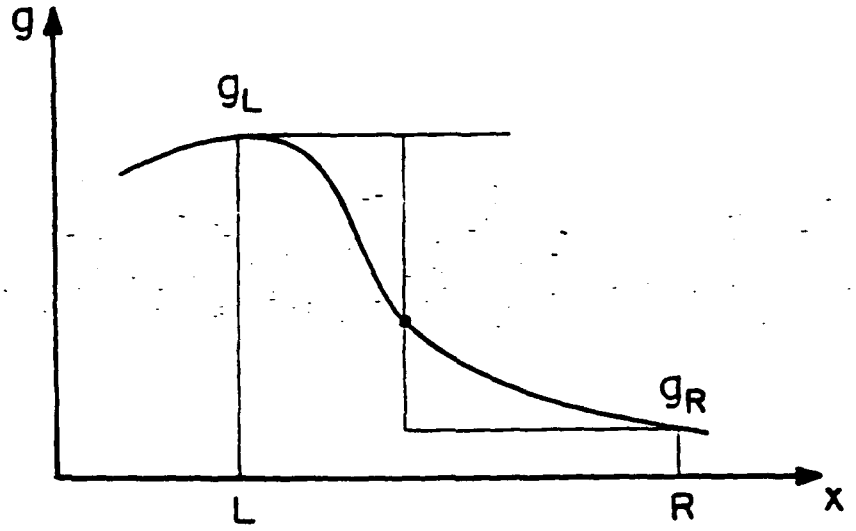

Figure 2.3: Discontinuity location recovering.

Since we can assume

$$\iint_{\omega} s\,dx\,dy = h^2 \sum_{\omega} s_{i,j} + O(h^2), \tag{2.16}$$

we can deduce from (2.15),(2.14) and (2.12) that

$$\int_{\partial\omega} (f,g)_n\, d(\partial\omega) = h \sum_{\partial\omega} (f^h, g^h)_n + O(h^2). \tag{2.17}$$

But since there is $O(h^2)$ pointwise convergence along the lines $B_R B_L$,$B_L L$ and $R B_R$, on the remaining line LR we must have

$$\int_{LR} (f,g)_n\, d(\partial\omega) - h \sum_{LR} (f^h, g^h)_n = O(h^2), \tag{2.18}$$

despite the lack of pointwise second order convergence along this segment LR.

2.3.2 Implementation

We can assume without loss of generality that the segment LR is perpendicular to the y-direction. Then, the flux normal to this segment will be just $g(u)$.

The numerical fluxes values g_L and g_R are assumed to approximate the fluxes of the differential solution at the respective points with an accuracy of h^2. The values at points between L and R do not in general approximate the differential solution. Their only role is to indicate the discontinuity location.

Let us reconstruct fluxes of a discontinuous solution from the numerical solution we have. For this purpose we have to substitute the values in the transition layer by values, which approximate the differential solution. Therefore, we have to extrapolate them

from the grid points to the left of L and to the right of R into the segment LR (see Fig. 2.3). The values, extrapolated from the left will approximate fluxes of the differential solution on the left side of the discontinuity, and values extrapolated from the right will aproximate fluxes from the right side of the discontinuity. We just have to find the proper location of the discontinuity (on this segment LR). Denote the reconstructed fluxes g^h_{rec}. We shall locate the discontinuity in such a way, that the following relation will hold

$$\int_{LR} g^h_{rec}\, dx = h \sum_{LR} g^h. \tag{2.19}$$

This relation defines the discontinuity location to within $O(h^2)$ accuracy. Indeed, suppose first that the width of the transition layer behaves like $O(h)$. This means that the number of grid points involved in the transition layer remains the same on each grid. Then, in order to obtain h^2 accuracy in the discontinuity location it is enough to use just an extrapolation by constants from points L and R. In case the width of the transition layer behave like $O(h^{\frac{1}{2}})$, it is enough to use a linear extrapolation.

The argument presented here has important implications for multigrid methods. These implications are explained in Chap.3.

Chapter 3

Multigrid methods

In this chapter we first give a brief description of the multigrid solver for an elliptic problem. Then we show how to adapt a multigrid algorithm for a non-elliptic problem.

3.1 Elliptic case

3.1.1 Relaxation

Consider a difference scheme

$$L^h u^h = s^h \tag{3.1}$$

with certain boundary conditions, approximating a boundary-value problem for an elliptic differential equation $Lu = s$.

Suppose these difference equations are being solved by a certain relaxation performed in a certain ordering. The error after n relaxation sweeps is

$$v_n^h = u^h - u_n^h, \tag{3.2}$$

where u_n^h is the current solution approximation.

It can be observed in numerical experiments that convergence of the relaxation is fast in the beginning, but becomes very slow after few sweeps. This is because the relaxation appears to be very efficient in reducing of the non-smooth error components. When the error is smooth the convergence is slow. That is relaxation smoothes the error. However, the smooth error can be approximated on a coarser grid. This is the main idea of the multigrid methods.

3.1.2 Coarse grid correction

Assume the Eq.(3.1) to be linear. Let $\tilde{u}^h$ be the approximation obtained by a few relaxation sweeps on the fine grid. The residuals of the fine grid equations are then given by

$$r^h = s^h - L^h \tilde{u}^h \tag{3.3}$$

The error $v^h = u^h - \tilde{u}^h$ will satisfy

$$L^h v^h = r^h. \tag{3.4}$$

Since v^h is smooth it can be approximated by a coarse grid function v^H satisfying the equation

$$L^H v^H = r^H, \tag{3.5}$$

where L^H is a grid H difference approximation to the differential operator and

$$r^H = I_h^H r^h, \tag{3.6}$$

where I_h^H is a fine-to-coarse transfer operator (injection or a certain type of weighting).

After obtaining a certain approximate solution v^H of (3.5) we can use it to correct the fine grid solution in the following way

$$\tilde{u}^h \leftarrow \tilde{u}^h + I_H^h v^h, \tag{3.7}$$

where I_H^h is an interpolation operator.

The process of calculating r^h, transfering it to the coarse grid, solving the coarse grid equations for v^H and interpolating the result and adding it to the fine grid approximation is called "coarse grid correction".

3.1.3 Full approximation scheme

We have described a way to construct coarse grid equations for the error v^h in the fine grid approximation $\tilde{u}^h$ to the solution of (3.1). It is called the Correction Scheme (CS). Another way to do it is called the Full Aproximation Scheme (FAS). Coarse grid equations are written in terms of a different function: instead of v^H we use

$$u^H = \bar{I}_h^H \tilde{u}^h + v^H. \tag{3.8}$$

This coarse grid function approximates the full solution on the coarse grid. The equation for it is

$$L^H u^H = \bar{s}^H, \tag{3.9}$$

where

$$\bar{s}^H = L^H \bar{I}_h^H \tilde{u}^h + I_h^H r^h. \tag{3.10}$$

After solving this equation approximately, we use $\tilde{u}^H$ to correct the fine grid approximation in the following way

$$\tilde{u}^h \leftarrow \tilde{u}^h + I_H^h(\tilde{u}^H - \bar{I}_h^H \tilde{u}^h). \tag{3.11}$$

FAS is used in case of nonlinear problem or when local refinement is needed.

Both CS and FAS schemes can be applied recursively. Such a solution method will be very efficient not only because coarser grid consist of less grid points. The error components, which are smooth on the finer grid, will look "less" smooth on coarser grids and therefore can be efficiently reduced there by relaxation.

3.1.4 Multigrid cycle

Assume we have a sequence of grids $h_1 > h_2 > \ldots > h_M$, where $h_M \equiv h$ and $h_k = 2h_{k+1}$. The k grid equation is written as

$$L^k u^k = s^k. \tag{3.12}$$

For $k = M$: $L^k = L^h$ and $s^k = s^h$. In the case of the Correction Scheme

$$s^{k-1} = I_k^{k-1}(s^k - L^k \tilde{u}^k) \tag{3.13}$$

and the initial approximation for u^{k-1} is zero. For the Full Aproximation Scheme

$$s^{k-1} = L^{k-1} \bar{I}_k^{k-1} \tilde{u}^k + I_k^{k-1}(s^k - L^k \tilde{u}^k) \tag{3.14}$$

and the initial approximation for u^{k-1} is $\bar{I}_k^{k-1} \tilde{u}_k$

Given an approximate solution $\tilde{u}^k$ of the Eq.(3.12). The multigrid cycle

$$\tilde{u}^k \leftarrow MG(k, \tilde{u}^k, s^k) \tag{3.15}$$

is defined recursively as follows:

- If $k = 1$, solve (3.12) by relaxations.
- Otherwise:
  - Perform ν_1 relaxation sweeps on (3.12) resulting in a new approximation.
  - Transfer the problem to the coarser grid $k-1$ and perform γ successive cycles

$$\tilde{u}^{k-1} \leftarrow MG(k-1, \tilde{u}^{k-1}, s^{k-1}). \tag{3.16}$$

  - Interpolate the correction to grid k and add it to the approximate solution there.
  - Perform ν_2 relaxation sweeps on grid k resulting in $\tilde{u}^k$ of (3.15).

When $\gamma = 1$ the corresponding multigrid cycle is called V cycle or $V(\nu_1, \nu_2)$. If $\gamma = 2$ – it is called W cycle or $W(\nu_1, \nu_2)$.

3.1.5 Full multigrid algorithm

The first approximation for the multigrid cycle on a certain level can be obtained by interpolating a solution from the previous coarser level, which itself has been calculated by multigrid cycles MG on that level and so on recursively. Such an algorithm is called a Full Multigrid Algorithm (FMG). We shall denote by $FMG(N, N_M, C, M)$ a certain FMG algorithm,where M is the finest level, N_M - the number of multigrid cycles performed in this finest level, N - number of cycles performed on the intermediate levels and C is the type of a multigrid cycle employed ($V(\nu_1, \nu_2)$ or $W(\nu_1, \nu_2)$). In case we want only to clarify how many MG cycles are done on the finest level we shall use the notation $N_M - FMG$.

3.2 Non-elliptic case

The first task in this case is to construct a stable discretization, such that a certain non-expensive local relaxation can be applied with it and will have good smoothing properties. It is sufficient for this purpose if the discrete equations are h-elliptic (see [2, 3]), which can be achieved by adding artificial viscosity with the coefficient proportional to h. Of course, this restricts us to the first order accuracy, but once it is done, the usual multigrid algorithms can be efficiently applied in case of a smooth solution. However, the discontinuous cases and the possibility to obtain second order accuracy require further studies. We require from our discretization to be stable, second order accurate and to provide a good resolution of discontinuities (representing them by thin and monotonic transition layers). Chapters 4,5 are devoted to the construction of such discretization schemes.

Since the difference schemes, which will be developed in Chapters 4,5 are nonlinear even in case of a linear differential equation, the Full Approximation Scheme should be used. We have shown in Chap.2 what should be understood by a discontinuity location when a shock capturing approach is used. The question is now how to perform coarsening in the multigrid process in order to achieve the same efficiency in converging the discontinuity to its correct location as in converging the solution in the smooth regions.

We have demonstrated that the conservation property of the difference scheme is of crucial importance for obtaining the correct discontinuity location. Therefore, in order to obtain a proper coarse grid correction for the discontinuity location a conservative residual transfer should also be used. This can be just the usual Full-Weighting.

Another question is what correction interpolation should be used for this purpose. The important consideration here is preserving the flux correction integral along coarse grid computational cells boundaries. Since this correction is small (when an FMG algorithm is used it is $O(h^2)$ in smooth regions and $O(h)$ in the neighborhood of a discontinuity), it is enough to preserve integral of the solution correction itself. This is perfectly done by the usual bilinear interpolation, which was shown already to perform well in smooth regions. Therefore, the conclusion is that for obtaining a full multigrid efficiency for converging the discontinuity location the usual coarsening techniques can be used. In other words, the coarse grid correction is perfectly capable of moving the discontinuity, commensurably with the solution changes it affects. There is no need to freeze the discontinuity location before going to coarse grids and performing a special procedure to update it after coming back to the fine grid (like it was suggested in [5]).

Chapter 4

Discretization: Linear case

The problem we shall deal with now is how to construct numerical fluxes

$$f_{i+\frac{1}{2},j},\ f_{i-\frac{1}{2},j},\ g_{i,j+\frac{1}{2}},\ g_{i,j-\frac{1}{2}}.$$

The general nonlinear case will be considered in the next chapter. In the present chapter we shall study some existing approaches as well as our own on the case of the simple linear constant coefficient equation

$$-\epsilon\Delta u + au_x + bu_y = s, \tag{4.1}$$

where $s = s(x, y)$. Without loss of generality we can assume

$$b \geq a \geq 0,\ b > 0. \tag{4.2}$$

Since the fluxes $f_{i+\frac{1}{2},j}$ and $f_{i-\frac{1}{2},j}$ are constructed in the same way (as well as $g_{i,j+\frac{1}{2}}$ and $g_{i,j-\frac{1}{2}}$) we shall give formulas for $f_{i-\frac{1}{2},j}$ and $g_{i,j-\frac{1}{2}}$ only.

4.1 Some existing methods

4.1.1 Central differencing

The most straightforward approach is the following "central" fluxes:

$$\begin{aligned} f^c{}_{i-\frac{1}{2},j} &= \tfrac{1}{2}a(u_{i,j} + u_{i-1,j}) \\ g^c{}_{i,j-\frac{1}{2}} &= \tfrac{1}{2}b(u_{i,j} + u_{i,j-1}). \end{aligned} \tag{4.3}$$

When (4.3) is substituted into the balance equation (2.11), it will give the standard central 5-point "star" second order accurate approximation to the equation (4.1)

$$\frac{1}{h}\left(a\frac{u_{i+1,j} - u_{i-1,j}}{2} + b\frac{u_{i,j+1} - u_{i,j-1}}{2}\right) = s_{i,j}. \tag{4.4}$$

This scheme does not have a good measure of ellipticity (see [3]). There exist certain high-frequency components which can be present in the error, but do not express themselves in residuals.

4.1.2 Upstream differencing

Let us add additional terms to the fluxes defined by (4.3) to produce "upstream" fluxes:

$$\begin{aligned} f^u{}_{i-\frac{1}{2},j} &= f^c{}_{i-\frac{1}{2},j} - \tfrac{1}{2}a(u_{i,j} - u_{i-1,j}) = au_{i-1,j} \\ g^u{}_{i,j-\frac{1}{2}} &= g^c{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}b(u_{i,j} - u_{i,j-1}) = bu_{i,j-1}. \end{aligned} \tag{4.5}$$

These additional terms correspond to the following anisotropic artificial viscosity

$$-\frac{1}{2}h((au_x)_x + (bu_y)_y), \tag{4.6}$$

which has the same sign as the physical viscosity. When fluxes (4.5) are substituted into the balance equation, we obtain a first order accurate upstream scheme, based on a 3-point stencil which can also be written in the form

$$u_{i,j} = \frac{a}{a+b}u_{i-1,j} + \frac{b}{a+b}u_{i,j-1}. \tag{4.7}$$

Since the coefficients in the right-hand side of this equation are positive (i.e. this difference operator is of the positive type), the following relation will hold:

$$\min(u_{i-1,j}, u_{i,j-1}) \le u_{i,j} \le \max(u_{i-1,j}, u_{i,j-1}). \tag{4.8}$$

In other words, a certain maximum principle holds for the solution of (4.5). We can summarize that the addition of the artificial viscosity to the central second order scheme creates a stable upstream first order scheme which produces a monotonic solution. The characteristic components (i.e., components, which are smooth in the characteristic direction and oscillating otherwise) and discontinuities propagating in an oblique direction will be smeared significantly in the solution, but those propagating in the grid direction (i.e., when $a = 0$) will be resolved perfectly.

4.1.3 Upstream second order scheme

We would like to construct a second order accurate scheme, which will maintain the stability property of (4.5). We can approximate derivatives in each direction using second order accurate 3-point one-sided approximation. The difference equation will be

$$a\frac{3u_{i,j} - 4u_{i-1,j} + u_{i-2,j}}{2h} + b\frac{3u_{i,j} - 4u_{i,j-1} + u_{i,j-2}}{2h} = s_{i,j}. \tag{4.9}$$

In terms of numerical fluxes it can be written

$$\begin{aligned} f^{u2}{}_{i-\frac{1}{2},j} &= f^u{}_{i-\frac{1}{2},j} + \tfrac{1}{2}a(u_{i-1,j} - u_{i-2,j}) \\ g^{u2}{}_{i,j-\frac{1}{2}} &= g^u{}_{i,j-\frac{1}{2}} + \tfrac{1}{2}b(u_{i,j-1} - u_{i,j-2}), \end{aligned} \tag{4.10}$$

The role of the additional terms is to compensate for the loss of accuracy due to the artificial viscosity (diffusion). Therefore, they are often called antidiffusive fluxes. However, when applied in the neighborhood of a discontinuity, this scheme will produce spurious oscillations in the solution.

4.1.4 Remark on overshoots

Consider a first order scheme. Its truncation error is dominated by its artificial viscosity, which behaves like a physical one. Therefore, the truncation error causes at least as much second order dissipation at every point as the physical viscosity. As a result, the difference solution has a monotonicity property analogous to one of the differential solution and overshoots do not appear. In case of the higher order difference scheme, its truncation error is dominated by higher order derivatives and does not behave like a physical viscosity anymore. It may even attain values which are opposite in sign to that of the second order dissipative term at some grid points. This means that some non-physical amounts of preserved quantity (mass, momentum or energy) are injected at these points. In the neighborhood of discontinuities, where derivatives become large, these local violations of the conservation law cause appearance of large overshoots. However, if the difference scheme is conservative, the conservation law still holds globally.

The usual way to overcome this difficulty is to multiply antidiffusive fluxes by a certain quantity, which is called a limiter. The concrete values of the limiter at each grid point should be chosen from considerations of the second order accuracy and monotonicity. It is closed to 1 in the smooth regions, not distorting the order of accuracy of the original scheme. However, it becomes different from 1 in the regions of changing gradients: introducing first order artificial viscosity needed to damp the oscillations if smaller than 1, or sharpening transition layers representing discontinuities (artificial compression) when larger than 1. This approach appeared to be very successful for 1D problems (the high resolution schemes are based on this principle, see [11, 12, 14, 16]). However, a straightforward extension for 2D (dimensional splitting) has some flaws. It leads to wide stencils and may require complicated block-relaxation process in order to maintain its stability.

4.1.5 Approach of Spekreijse

Let us multiply the antidiffusive fluxes in (4.10) by a certain limiter $\psi(R)$, producing:

$$\begin{aligned} f^a{}_{i-\frac{1}{2},j} &= f^u{}_{i-\frac{1}{2},j} + \tfrac{1}{2}\psi(R^a{}_{i-\frac{1}{2},j})a(u_{i-1,j} - u_{i-2,j}) \\ g^a{}_{i,j-\frac{1}{2}} &= g^u{}_{i,j-\frac{1}{2}} + \tfrac{1}{2}\psi(Q^a{}_{i,j-\frac{1}{2}})b(u_{i,j-1} - u_{i,j-2}), \end{aligned} \tag{4.11}$$

where

$$R^a{}_{i-\frac{1}{2},j} = \frac{u_{i,j} - u_{i-1,j}}{u_{i-1,j} - u_{i-2,j}} \tag{4.12}$$

$$Q^a{}_{i,j-\frac{1}{2}} = \frac{u_{i,j} - u_{i,j-1}}{u_{i,j-1} - u_{i,j-2}}. \tag{4.13}$$

Choose ψ to be the Van Albada's limiter

$$\psi_{VA}(R) = \frac{R^2 + R}{R^2 + 1}. \tag{4.14}$$

It is shown in [20] that the solution of (4.11) will have the same monotonic property (4.8) as the solution of (4.5), and will be second order accurate in the smooth regions.

However, a pointwise relaxation will be unstable when applied to this scheme. The only way to maintain the stability is to use more expensive block relaxation (4-points relaxation, [20]). This scheme can be regarded as a typical example of the dimensional splitting approach.

4.2 2D approach: Homogeneous equation

Consider the Eq.(4.1) with $s(x,y) \equiv 0$.

4.2.1 2D scheme

Our approach is to construct upstream antidiffusive fluxes using certain difference approximations to the original equation (4.1) itself. It leads to a "compact" stencil, which will use diagonal grid points instead of the points which are distance $2h$ from the central one. The difference scheme based on such a stencil will have a smaller truncation error and will be applicable near the boundaries as well. Another advantage is the possibility to separate between the treatment of the cross-stream and streamwise directions. This allows us to add the high resolution in the cross-stream direction only, but to maintain a good measure of ellipticity in the streamwise direction. As a result we can obtain a discrete scheme which is able to resolve well characteristic components (see Chap.2) and discontinuities together with good stability properties.

Define the following 2D compact scheme

$$\begin{aligned} f^{2D}{}_{i-\frac{1}{2},j} &= f^{u}{}_{i-\frac{1}{2},j} - \tfrac{1}{2}b(u_{i-1,j} - u_{i-1,j-1}) \\ g^{2D}{}_{i,j-\frac{1}{2}} &= g^{u}{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}a(u_{i,j-1} - u_{i-1,j-1}) \end{aligned} \tag{4.15}$$

Lemma 4.1 *Scheme* (4.15) *is second order accurate.*

Proof: We want to show that

$$\begin{aligned} f^{2D}{}_{i+\frac{1}{2},j} - f^{2D}{}_{i-\frac{1}{2},j} &= h(au_x) + O(h^3) \\ g^{2D}{}_{i,j+\frac{1}{2}} - g^{2D}{}_{i,j-\frac{1}{2}} &= h(bu_y) + O(h^3). \end{aligned} \tag{4.16}$$

Rewrite (4.15) as

$$\begin{aligned} f^{2D}{}_{i-\frac{1}{2},j} &= f^{c}{}_{i-\frac{1}{2},j} - \tfrac{1}{2}(a(u_{i,j} - u_{i-1,j}) + b(u_{i-1,j} - u_{i-1,j-1})) \\ g^{2D}{}_{i,j-\frac{1}{2}} &= g^{c}{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}(b(u_{i,j} - u_{i,j-1}) + a(u_{i,j-1} - u_{i-1,j-1})) \end{aligned} \tag{4.17}$$

Note, that

$$\begin{aligned} f^{2D}{}_{i+\frac{1}{2},j} - f^{2D}{}_{i-\frac{1}{2},j} &= f^{c}{}_{i+\frac{1}{2},j} - f^{c}{}_{i-\frac{1}{2},j} - \\ \frac{1}{2}(a(u_{i+1,j} - u_{i,j}) &+ b(u_{i,j} - u_{i,j-1})) + \\ \frac{1}{2}(a(u_{i,j} - u_{i-1,j}) &+ b(u_{i-1,j} - u_{i-1,j-1})) \end{aligned} \tag{4.18}$$

Since the central scheme (4.3) is second order accurate, we can conclude that

$$\begin{aligned} f^{2D}{}_{i+\frac{1}{2},j} &- f^{2D}{}_{i-\frac{1}{2},j} = \\ h(au_x) &- \frac{1}{2}h^2(au_x + bu_y)_x + O(h^3), \end{aligned} \tag{4.19}$$

or taking the equation (4.1) into account

$$\begin{aligned} f^{2D}{}_{i+\frac{1}{2},j} &- f^{2D}{}_{i-\frac{1}{2},j} = \\ h(au_x) &+ O(h^3), \end{aligned} \tag{4.20}$$

Similarly

$$\begin{aligned} g^{2D}{}_{i,j+\frac{1}{2}} - g^{2D}{}_{i,j-\frac{1}{2}} = h(bu_y) &- \frac{1}{2}h^2(au_x + bu_y)_y + O(h^3) = \\ h(bu_y) &+ O(h^3). \end{aligned} \tag{4.21}$$

Remark 4.1 There is no first order artificial viscosity used by the scheme (4.15).

When (4.15) are substituted into the balance equation they give:

$$u_{i,j} = u_{i-1,j-1} + \frac{b-a}{a+b}u_{i,j-1} + \frac{a-b}{a+b}u_{i-1,j}. \tag{4.22}$$

Note that the diagonal grid point value $u_{i-1,j-1}$ participates in the difference equation instead of the values $u_{i-2,j}$ and $u_{i,j-2}$ in (4.11). This scheme is second order accurate, but not of the positive type, because the coefficient of $u_{i-1,j}$ is negative.

Remark 4.2 Another significant defect of this scheme is that in the case $a \to 0$ it leads to the difference equation

$$u_{i,j} = u_{i-1,j-1} + (u_{i,j-1} - u_{i-1,j}) \tag{4.23}$$

which is based on the wide stencil. However, it is natural to expect the following equation in this case

$$u_{i,j} = u_{i-1,j-1}. \tag{4.24}$$

This defect creates an obvious difficulty for extending this approach for the case of the variable coefficient equation and for the non-linear case as well. This is because it may lead to discontinuous numerical fluxes.

4.2.2 First order N scheme

Modify the previous scheme:

$$\begin{aligned} f^o_{i-\frac{1}{2},j} &= au_{i-1,j} - \tfrac{1}{2}\min(a,b)(u_{i-1,j} - u_{i-1,j-1}) \\ g^o_{i,j-\frac{1}{2}} &= bu_{i,j-1} - \tfrac{1}{2}\min(a,b)(u_{i,j-1} - u_{i-1,j-1}) \end{aligned} \tag{4.25}$$

In our case (4.2), (4.25) can be rewritten as

$$\begin{aligned} f^o_{i-\frac{1}{2},j} &= au_{i-1,j} - \tfrac{1}{2}a(u_{i-1,j} - u_{i-1,j-1}) \\ g^o_{i,j-\frac{1}{2}} &= bu_{i,j-1} - \tfrac{1}{2}a(u_{i,j-1} - u_{i-1,j-1}). \end{aligned} \tag{4.26}$$

When substituted into the balance equation it will give

$$u_{i,j} = \frac{a}{b}u_{i-1,j-1} + \frac{b-a}{b}u_{i,j-1}. \tag{4.27}$$

This scheme is already of positive type, however only first order accurate. This *type* of scheme was actually suggested in [6, 7], however, the present formulation is much simpler (by avoiding rotation transformation of the equation) and is obviously conservative. We shall call it N scheme ("narrow" scheme). Its solution will satisfy the following monotonicity property:

$$\min(u_{i,j-1}, u_{i-1,j-1}) \le u_{i,j} \le \max(u_{i,j-1}, u_{i-1,j-1}). \tag{4.28}$$

Notice, that this scheme will perfectly resolve discontinuities that are aligned with a diagonal direction as well as along grid lines. It can be shown that the cross-stream viscosity coefficient of this scheme is at least 3.64 times smaller (see [18]) than that of (4.5). Still, the resolution of an oblique discontinuity by such a scheme has to be improved.

4.2.3 S scheme

We would like to correct the N scheme to be second order accurate in a way similar to (4.11), retaining the monotonic property (4.28). This can be done by the following S scheme:

$$\begin{aligned} f^S_{i-\frac{1}{2},j} &= f^o_{i-\frac{1}{2},j} - \tfrac{1}{2}\psi(R_{i-\frac{1}{2},j})(b-a)(u_{i-1,j} - u_{i-1,j-1}) \\ g^S_{i,j-\frac{1}{2}} &= g^o_{i,j-\frac{1}{2}}, \end{aligned} \tag{4.29}$$

where

$$R_{i-\frac{1}{2},j} = \frac{-a(u_{i,j-1} - u_{i-1,j-1})}{b(u_{i-1,j} - u_{i-1,j-1})}. \tag{4.30}$$

The question is what conditions have to be imposed on the limiter function $\psi(R)$, in order to ensure monotonicity and second order accuracy of the S scheme. The two following lemmas and the remark answer this question and show that all the limiters used in 1D problems and reviewed in [21] are good for this purpose.

Lemma 4.2 *If the limiter $\psi = \psi(R)$ satisfies the following inequality*

$$0 \leq \frac{\psi(R)}{R} \leq 2, \ \psi(R) \leq 2, \tag{4.31}$$

then the monotonicity property (4.28) *holds for the S scheme.*

Proof: Rewrite (4.30) as

$$u_{i-1,j} - u_{i-1,j-1} \equiv -\frac{1}{R_{i-\frac{1}{2},j}} \frac{a}{b} (u_{i,j-1} - u_{i-1,j-1}). \tag{4.32}$$

Then the correction for $f^{\circ}_{i-\frac{1}{2},j}$ in (4.29) can be replaced by

$$\frac{1}{2}\psi(R_{i-\frac{1}{2},j})(b-a)(u_{i-1,j} - u_{i-1,j-1}) \equiv -\frac{1}{2}\frac{\psi(R_{i-\frac{1}{2},j})}{R_{i-\frac{1}{2},j}} \frac{a(b-a)}{b} (u_{i,j-1} - u_{i-1,j-1}) \tag{4.33}$$

When the numerical fluxes (defined by (4.29)) and the correction for the flux $f^{\circ}_{i-\frac{1}{2},j}$ as calculated using (4.33) are substituted into the balance equation, we obtain the following relation:

$$\begin{aligned} & a(1 - \tfrac{1}{2}\tfrac{\psi(R_{i-\frac{1}{2},j})}{R_{i-\frac{1}{2},j}}(1 - \tfrac{a}{b})) \ \ (u_{i,j} - u_{i-1,j-1}) \\ +(b-a)(1 - \tfrac{1}{2}\psi(R_{i+\frac{1}{2},j}) & + \tfrac{1}{2}\tfrac{\psi(R_{i-\frac{1}{2},j})}{R_{i-\frac{1}{2},j}}\tfrac{a}{b}) \ \ (u_{i,j} - u_{i,j-1}) = 0. \end{aligned} \tag{4.34}$$

It is easy to see that condition (4.31) ensures the positiveness of the coefficients in (4.34), and therefore, monotonicity of the S scheme.

Remark 4.3 It is easy to see from (4.34) that the S scheme in case $a \to 0$ leads to the difference equation based upon the narrow stencil

$$u_{i,j} = u_{i,j-1}. \tag{4.35}$$

This means that the S scheme does not suffer from the same defect as the 2D scheme.

Lemma 4.3 *If $\psi = \psi(R) \in C^2$ and*

$$\psi(1) = 1, \tag{4.36}$$

then the S scheme is second order accurate.

Proof: In order to prove this lemma it is enough to show that

$$f^{S}_{i+\frac{1}{2},j} - f^{S}{}_{i-\frac{1}{2},j} = h(au_x) + O(h^3). \tag{4.37}$$

$$f^{2D}_{i-\frac{1}{2},j} - f^{S}{}_{i-\frac{1}{2},j} = \frac{1}{2}(1-\psi(R_{i-\frac{1}{2},j}))(b-a)(u_{i-1,j}-u_{i-1,j-1}) \tag{4.38}$$

Taking (4.36) into account:

$$1-\psi(R_{i-\frac{1}{2},j}) = -\psi'_R(1)(R_{i-\frac{1}{2},j}-1)+O(h^2), \tag{4.39}$$

and hence

$$\begin{aligned} f^{2D}_{i-\frac{1}{2},j} - f^{S}{}_{i-\frac{1}{2},j} &= -\frac{1}{2}\psi'_R(1)(R_{i-\frac{1}{2},j}-1)(b-a)(u_{i-1,j}-u_{i-1,j-1})+O(h^3) \\ &= \psi'_R(1)\frac{b-a}{b}(b(u_{i-1,j}-u_{i-1,j-1})+a(u_{i,j-1}-u_{i-1,j-1}))+O(h^3). \end{aligned} \tag{4.40}$$

Therefore

$$(f^{2D}_{i+\frac{1}{2},j} - f^{S}{}_{i+\frac{1}{2},j}) - (f^{2D}_{i-\frac{1}{2},j} - f^{S}{}_{i-\frac{1}{2},j}) = O(h^3) \tag{4.41}$$

or

$$f^{S}_{i+\frac{1}{2},j} - f^{S}{}_{i-\frac{1}{2},j} = f^{2D}_{i+\frac{1}{2},j} - f^{2D}{}_{i-\frac{1}{2},j}+O(h^3), \tag{4.42}$$

which proves (4.37).

Remark 4.4 Suppose that the limiter $\psi(R)$ is twice continuously differentiable only in the neighborhood of $R=1$ and Lipschitz continuous otherwise. It is clear from the proof of the previous lemma that the S scheme employing such a limiter is also second order accurate. Moreover, even if ψ'_R is discontinuous at $R=1$, the S scheme will be still second order accurate in L_1 norm. This is because the order of approximation will deteriorate to the first due to discontinuities in the derivative of the limiter in such a case only in computational cells located along isolated characteristic lines. Therefore, these computational cells will cover only $O(h)$ part of the domain.

Remark 4.5 Note that a limiter satisfying both (4.31) and (4.36) cannot be a linear function. Therefore, a monotone higher order accurate scheme will be nonlinear even in case of a linear equation.

4.2.4 Examples of limiters

The construction of a limiter is not unique. Therefore, several different limiters were proposed in the 1D framework (see [21]) and can serve our purpose. The difference between different limiters is only in the amount of artificial viscosity they may add to the scheme in order to damp oscillations and in the amount of artificial compression they may or may not add to the scheme.

We shall give here some examples of limiters.

Example 4.1 Van Leer limiter

$$\psi_{VL} = \frac{|R|+R}{1+|R|} \tag{4.43}$$

Note that $\psi_{VL} \geq 0$ is a monotone increasing function

$$\psi_{VL} = \begin{cases} 0, & \text{if } R < 0 \\ \frac{2R}{1+R}, & \text{if } R \geq 0. \end{cases} \tag{4.44}$$

Example 4.2 A class of limiters defined by

$$\psi_\phi = \max(0, \min(\phi R, 1), \min(R, \phi)), \tag{4.45}$$

where $1 \leq \phi \leq 2$.

Note that ψ_ϕ is a monotone increasing function. When $\phi = 2$, $\psi_\phi(R)$ corresponds to the Roe highly compressive "superbee" limiter, defined as

$$\psi_2 = \max(0, \min(2R, 1), \min(R, 2)). \tag{4.46}$$

This limiter introduces a large artificial compression where possible. This means that discontinuities in the solution will be resolved nicely, however, non-physical sharp layers can appear.

When $\phi = 1$, $\psi_\phi(R)$ corresponds to another Roe limiter:

$$\psi_1 = \max(0, \min(R, 1)). \tag{4.47}$$

Note that $0 \leq \psi_1(R) \leq 1$. This means, that this limiter may add only an artificial viscosity when it is needed to damp the oscillations, but does not add any artificial compression. The solution produced by S scheme using this limiter will contain only layers representing the physical discontinuities. However, these layers will not be as sharp as in the previous case.

Remark 4.6 Van Albada limiter does not possess the property (4.31), therefore if used with the S scheme, a non-monotone solution may be produced. It's unique property is

$$\psi_{VA} \in C^\infty. \tag{4.48}$$

However, it is not needed for a second order accuracy. We shall also show that this property is not necessary to obtain a fast convergence of the multigrid algorithm.

4.2.5 Relaxation

Denote

$$F(u_{i,j}) \equiv h(f_{i+\frac{1}{2},j} - f_{i-\frac{1}{2},j} + g_{i,j+\frac{1}{2}} - g_{i,j-\frac{1}{2}}) = 0. \tag{4.49}$$

A pointwise relaxation sweep implies updating the value of the numerical solution $u_{i,j}$ at each internal grid point. This can be done performing one Newton iteration for the nonlinear equation (4.49)

$$u_{i,j}^{new} = u_{i,j}^{old} - \frac{F(u_{i,j}^{old})}{F'(u_{i,j}^{old})}. \tag{4.50}$$

To implement this formula, we have to differentiate with respect to $u_{i,j}$ all numerical fluxes participating in the balance equation. For the first order accurate upstream scheme these derivatives are

$$\begin{aligned} (f^u{}_{i+\frac{1}{2},j})'_{u_{i,j}} &= a \\ (f^u{}_{i-\frac{1}{2},j})'_{u_{i,j}} &= 0 \\ (g^u{}_{i,j+\frac{1}{2}})'_{u_{i,j}} &= b \\ (g^u{}_{i,j-\frac{1}{2}})'_{u_{i,j}} &= 0. \end{aligned} \tag{4.51}$$

For the first order N scheme:

$$\begin{aligned} (f^o{}_{i+\frac{1}{2},j})'_{u_{i,j}} &= \tfrac{1}{2}a \\ (f^o{}_{i-\frac{1}{2},j})'_{u_{i,j}} &= 0 \\ (g^o{}_{i,j+\frac{1}{2}})'_{u_{i,j}} &= b - \tfrac{1}{2}a \\ (g^o{}_{i,j-\frac{1}{2}})'_{u_{i,j}} &= 0. \end{aligned} \tag{4.52}$$

If we assume differentiability of the limiter ψ, the flux derivatives for the S schemes also appear to be as simple as for (4.11) (see [20]). They are:

$$\begin{aligned} (f^S{}_{i+\frac{1}{2},j})'_{u_{i,j}} &= (f^o{}_{i+\frac{1}{2},j})'_{u_{i,j}} - \tfrac{1}{2}(b-a)(\psi(R_{i+\frac{1}{2},j}) - \psi'_R(R_{i+\frac{1}{2},j})R_{i+\frac{1}{2},j}) \\ (f^S{}_{i-\frac{1}{2},j})'_{u_{i,j}} &= (f^o{}_{i-\frac{1}{2},j})'_{u_{i,j}} \\ (g^S{}_{i,j+\frac{1}{2}})'_{u_{i,j}} &= (g^o{}_{i,j+\frac{1}{2}})'_{u_{i,j}} \\ (g^S{}_{i,j-\frac{1}{2}})'_{u_{i,j}} &= (g^o{}_{i,j-\frac{1}{2}})'_{u_{i,j}} \end{aligned} \tag{4.53}$$

Newton's method is known to have a quadratic convergence, while the relaxation process is only linearly convergent, even locally. Therefore, relaxation does not really take advantage of the fast convergence of the Newton iterations. This is the reason why the flux derivatives can be calculated approximately, treating the limiter function as a quantity independent of $u_{i,j}$. Moreover, the flux derivatives of the S scheme can be substituted by the flux derivatives of the N scheme. The numerical experiments confirm that this substitution does not influence the performance of the multigrid solver.

We can conclude that it is not neccessary for limiters to be C^2 functions for the fast multigrid convergence as well as for second order accuracy.

4.3 2D approach: Inhomogeneous equation

The S scheme, as constructed previously, is second order accurate in case of a homogeneous equation. We shall generalize the S scheme in order to maintain the second order accuracy for an inhomogeneous problem (4.1).

4.3.1 2D scheme

Denote

$$\begin{aligned} s_{i-\frac{1}{2},j} &= \tfrac{1}{2}(s_{i,j} + s_{i-1,j}) \\ s_{i,j-\frac{1}{2}} &= \tfrac{1}{2}(s_{i,j} + s_{i,j-1}) \end{aligned} \tag{4.54}$$

Define

$$\begin{aligned} f^{2D}{}_{i-\frac{1}{2},j} &= f^{c}{}_{i-\frac{1}{2},j} - \tfrac{1}{2}(a(u_{i,j} - u_{i-1,j}) + b(u_{i-1,j} - u_{i-1,j-1}) - hs_{i-\frac{1}{2},j}) \\ g^{2D}{}_{i,j-\frac{1}{2}} &= g^{c}{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}(b(u_{i,j} - u_{i,j-1}) + a(u_{i,j-1} - u_{i-1,j-1}) - hs_{i,j-\frac{1}{2}}) \end{aligned} \tag{4.55}$$

Lemma 4.4 *Scheme* (4.55) *is second order accurate.*

Proof: This lemma can be proved in a similar way to the corresponding homogeneous case. We want to show that

$$\begin{aligned} f^{2D}{}_{i+\frac{1}{2},j} - f^{2D}{}_{i-\frac{1}{2},j} &= h(au_x) + O(h^3) \\ g^{2D}{}_{i,j+\frac{1}{2}} - g^{2D}{}_{i,j-\frac{1}{2}} &= h(bu_y) + O(h^3). \end{aligned} \tag{4.56}$$

Note, that

$$\begin{aligned} f^{2D}{}_{i+\frac{1}{2},j} - f^{2D}{}_{i-\frac{1}{2},j} &= f^{c}{}_{i+\frac{1}{2},j} - f^{c}{}_{i-\frac{1}{2},j} - \\ \frac{1}{2}(a(u_{i+1,j} - u_{i,j}) &+ b(u_{i,j} - u_{i,j-1}) - hs_{i+\frac{1}{2},j}) + \\ \frac{1}{2}(a(u_{i,j} - u_{i-1,j}) &+ b(u_{i-1,j} - u_{i-1,j-1}) - hs_{i-\frac{1}{2},j}) \end{aligned} \tag{4.57}$$

Since the central scheme (4.3) is second order accurate, we can conclude that

$$\begin{aligned} f^{2D}{}_{i+\frac{1}{2},j} &- f^{2D}{}_{i-\frac{1}{2},j} = \\ h(au_x) &- \frac{1}{2}h^2(au_x + bu_y - s)_x + O(h^3), \end{aligned} \tag{4.58}$$

or taking the entire equation into account

$$\begin{aligned} f^{2D}{}_{i+\frac{1}{2},j} &- f^{2D}{}_{i-\frac{1}{2},j} = \\ = h(au_x) &+ O(h^3). \end{aligned} \tag{4.59}$$

Similarly

$$\begin{aligned} g^{2D}{}_{i,j+\frac{1}{2}} - g^{2D}{}_{i,j-\frac{1}{2}} = h(bu_y) &- \frac{1}{2}h^2(au_x + bu_y - s)_y + O(h^3) = \\ h(bu_y) &+ O(h^3). \end{aligned} \tag{4.60}$$

4.3.2 N scheme

In case one is interested to obtain first order accuracy the previously defined N scheme can be used. However, the purpose of it here is to serve as an intermediate step towards the construction of the second order accurate S scheme. Therefore, we shall modify it. For our representative case $a < b$, putting

$$\beta = \min(1, \frac{a}{b}) = \frac{a}{b} < 1, \tag{4.61}$$

the N scheme can be defined by:

$$\begin{aligned} f^{\circ}{}_{i-\frac{1}{2},j} &= f^{c}{}_{i-\frac{1}{2},j} - \tfrac{1}{2}(a(u_{i,j} - u_{i-1,j}) + \beta(b(u_{i-1,j} - u_{i-1,j-1}) - hs_{i-\frac{1}{2},j})) \\ g^{\circ}{}_{i,j-\frac{1}{2}} &= g^{c}{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}(b(u_{i,j} - u_{i,j-1}) + a(u_{i,j-1} - u_{i-1,j-1}) - hs_{i,j-\frac{1}{2}}). \end{aligned} \tag{4.62}$$

4.3.3 S scheme

The S scheme is defined by

$$\begin{aligned} f^S{}_{i-\frac{1}{2}j} &= f^o{}_{i-\frac{1}{2}j} - \tfrac{1}{2}\psi(R_{i-\frac{1}{2}j})(1-\beta)(b(u_{i-1j} - u_{i-1j-1}) - hs_{i-\frac{1}{2}j}) \\ g^S{}_{ij-\frac{1}{2}} &= g^o{}_{ij-\frac{1}{2}} \end{aligned} \tag{4.63}$$

where

$$R_{i-\frac{1}{2}j} = \frac{-a(u_{ij-1} - u_{i-1j-1})}{b(u_{i-1j} - u_{i-1j-1}) - hs_{i-\frac{1}{2}j}} \tag{4.64}$$

Lemma 4.5 *If $\psi = \psi(R) \in C^2$ and*

$$\psi(1) = 1, \tag{4.65}$$

then the S scheme is second order accurate.

Proof: In order to prove this lemma it is sufficient to show that

$$f^S{}_{i+\frac{1}{2}j} - f^S{}_{i-\frac{1}{2}j} = h(au_x) + O(h^3). \tag{4.66}$$

Observe that

$$f^{2D}_{i-\frac{1}{2}j} - f^S{}_{i-\frac{1}{2}j} = \frac{1}{2}(1 - \psi(R_{i-\frac{1}{2}j}))(1-\beta)(b(u_{i-1j} - u_{i-1j-1}) - hs_{i-\frac{1}{2}j}) \tag{4.67}$$

Taking into account:

$$1 - \psi(R_{i-\frac{1}{2}j}) = -\psi'_R(1)(R_{i-\frac{1}{2}j} - 1) + O(h^2) \tag{4.68}$$

Then

$$\begin{aligned} f^{2D}_{i-\frac{1}{2}j} - f^S{}_{i-\frac{1}{2}j} &= \\ -\frac{1}{2}\psi'_R(1)(R_{i-\frac{1}{2}j} - 1)(1 &- \beta)(b(u_{i-1j} - u_{i-1j-1}) - hs_{i-\frac{1}{2}j}) + O(h^3) = \\ \psi'_R(1)\frac{b-a}{b}(b(u_{i-1j} &- u_{i-1j-1}) + a(u_{ij-1} - u_{i-1j-1}) + hs_{i-\frac{1}{2}j}) + O(h^3). \end{aligned}$$

Therefore

$$f^S{}_{i+\frac{1}{2}j} - f^S{}_{i-\frac{1}{2}j} = f^{2D}_{i+\frac{1}{2}j} - f^{2D}{}_{i-\frac{1}{2}j} + O(h^3), \tag{4.69}$$

which together with the fact that 2D scheme is second order accurate proves (4.66).

Remark 4.7 In order to obtain a second order accurate solution to the inhomogeneous problem by the *FMG* algorithm it is neccessary to use this scheme on the currently finest grid only. The scheme constructed for homogeneous problems can be employed on coarse grids. The coarse grid correction obtained this way will be only first order approximation to the needed correction on the finest grid. However, it is satisfactory, because the needed correction is only $O(h^2)$ large.

Chapter 5

Discretization: General case

There are several slightly different ways to extend the genuinely 2D approach for the case of a nonlinear conservation law. We shall present here one of the simplest. We shall first extend the construction of the difference scheme and prove its monotonicity for a homogeneous case. Then we shall treat an inhomogeneous case, demonstrating, that a second order accuracy can be obtained by this approach. A second order accuracy of the schemes constructed for a homogeneous case will follows directly from the more general result.

5.1 Homogeneous equation

Consider a 2D homogeneous conservation law

$$(f(u))_x + (g(u))_y = 0. \tag{5.1}$$

Denote

$$\begin{aligned} a(u) &= f_u \\ b(u) &= g_u \end{aligned} \tag{5.2}$$

Then (5.1) can be written in quasilinear form

$$a(u)u_x + b(u)u_y = 0. \tag{5.3}$$

5.1.1 Central and upstream schemes

A central unstable second order accurate difference approximation for (5.1) is determined by

$$\begin{aligned} f^c{}_{i-\frac{1}{2}j} &= \tfrac{1}{2}(f_{ij} + f_{i-1j}) \\ g^c{}_{ij-\frac{1}{2}} &= \tfrac{1}{2}(g_{ij} + g_{ij-1}). \end{aligned} \tag{5.4}$$

In order to stabilize (5.4), we add artificial viscosity term

$$\begin{aligned} f_{i-\frac{1}{2}j} &= f^c{}_{i-\frac{1}{2}j} - \tfrac{1}{2}|a_{i-\frac{1}{2}j}|(u_{ij} - u_{i-1j}) \\ g_{ij-\frac{1}{2}} &= g^c{}_{ij-\frac{1}{2}} - \tfrac{1}{2}|b_{ij-\frac{1}{2}}|(u_{ij} - u_{ij-1}). \end{aligned} \tag{5.5}$$

If

$$\begin{aligned} a_{i-\frac{1}{2},j} &= \frac{f(u_{i,j})-f(u_{i-1,j})}{u_{i,j}-u_{i-1,j}} \\ b_{i,j-\frac{1}{2}} &= \frac{g(u_{i,j})-g(u_{i,j-1})}{u_{i,j}-u_{i,j-1}}, \end{aligned} \tag{5.6}$$

then scheme (5.5) is exactly upstream and of positive type. However, relations (5.6) cannot be used for numerical calculations. Note, that for nonlinearities typical of fluid dynamics equations, the following is true

$$\begin{aligned} \tfrac{1}{2}(a(u_{i,j}) + a(u_{i-1,j})) &= \frac{f(u_{i,j})-f(u_{i-1,j})}{u_{i,j}-u_{i-1,j}} \\ \tfrac{1}{2}(b(u_{i,j}) + b(u_{i,j-1})) &= \frac{g(u_{i,j})-g(u_{i,j-1})}{u_{i,j}-u_{i,j-1}}, \end{aligned} \tag{5.7}$$

Therefore, the following definition will be used:

$$\begin{aligned} a_{i-\frac{1}{2},j} &= \tfrac{1}{2}(a(u_{i,j}) + a(u_{i-1,j})) \\ b_{i,j-\frac{1}{2}} &= \tfrac{1}{2}(b(u_{i,j}) + b(u_{i,j-1})). \end{aligned} \tag{5.8}$$

The resulting scheme will produce monotonic solutions, but only of first order accuracy.

5.1.2 2D scheme

Define

$$\begin{aligned} A_{i-\frac{1}{2},j-\frac{1}{2}} &= \tfrac{1}{2}(a(u_{i,j-\nu}) + a(u_{i-1,j-\nu})) \\ B_{i-\frac{1}{2},j-\frac{1}{2}} &= \tfrac{1}{2}(b(u_{i-\mu,j}) + b(u_{i-\mu,j-1})), \end{aligned} \tag{5.9}$$

where $\mu, \nu = 0, 1$, depending on our choice only. Denote

$$\begin{aligned} A^{+}_{i-\frac{1}{2},j-\frac{1}{2}} &= \max(0, A_{i-\frac{1}{2},j-\frac{1}{2}}) \\ A^{-}_{i-\frac{1}{2},j-\frac{1}{2}} &= \min(0, A_{i-\frac{1}{2},j-\frac{1}{2}}) \end{aligned} \tag{5.10}$$

and

$$\begin{aligned} B^{+}_{i-\frac{1}{2},j-\frac{1}{2}} &= \max(0, B_{i-\frac{1}{2},j-\frac{1}{2}}) \\ B^{-}_{i-\frac{1}{2},j-\frac{1}{2}} &= \min(0, B_{i-\frac{1}{2},j-\frac{1}{2}}). \end{aligned} \tag{5.11}$$

Denote also

$$(\delta_x u)_{i-\frac{1}{2},j-\frac{1}{2}} = \begin{cases} (u_{i,j} - u_{i-1,j}), & \text{if } B_{i-\frac{1}{2},j-\frac{1}{2}} \le 0 \\ (u_{i,j-1} - u_{i-1,j-1}), & \text{if } B_{i-\frac{1}{2},j-\frac{1}{2}} > 0 \end{cases} \tag{5.12}$$

$$(\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} = \begin{cases} (u_{i,j} - u_{i,j-1}), & \text{if } A_{i-\frac{1}{2},j-\frac{1}{2}} \le 0 \\ (u_{i-1,j} - u_{i-1,j-1}), & \text{if } A_{i-\frac{1}{2},j-\frac{1}{2}} > 0. \end{cases} \tag{5.13}$$

Then the genuinely 2D scheme can be given by

$$\begin{aligned} f^{2D}{}_{i-\frac{1}{2},j} &= f^{c}{}_{i-\frac{1}{2},j} - \tfrac{1}{2}(|a_{i-\frac{1}{2},j}|(u_{i,j} - u_{i-1,j}) + \phi^{+}_{i-\frac{1}{2},j} + \phi^{-}_{i-\frac{1}{2},j}) \\ g^{2D}{}_{i,j-\frac{1}{2}} &= g^{c}{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}(|b_{i,j-\frac{1}{2}}|(u_{i,j} - u_{i,j-1}) + \gamma^{+}_{i,j-\frac{1}{2}} + \gamma^{-}_{i,j-\frac{1}{2}}), \end{aligned} \tag{5.14}$$

where

$$\begin{aligned} \phi^{+}_{i-\frac{1}{2},j} &= \mathrm{sign}(A_{i-\frac{1}{2},j-\frac{1}{2}})B^{+}{}_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} \\ \phi^{-}_{i-\frac{1}{2},j} &= \mathrm{sign}(A_{i-\frac{1}{2},j+\frac{1}{2}})B^{-}{}_{i-\frac{1}{2},j+\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j+\frac{1}{2}} \end{aligned} \tag{5.15}$$

and

$$\begin{aligned} \gamma^{+}_{i,j-\frac{1}{2}} &= \mathrm{sign}(B_{i-\frac{1}{2},j-\frac{1}{2}})A^{+}{}_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2},j-\frac{1}{2}} \\ \gamma^{-}_{i,j-\frac{1}{2}} &= \mathrm{sign}(B_{i+\frac{1}{2},j-\frac{1}{2}})A^{-}{}_{i+\frac{1}{2},j-\frac{1}{2}}(\delta_x u)_{i+\frac{1}{2},j-\frac{1}{2}}. \end{aligned} \tag{5.16}$$

The second order accuracy of this scheme will be a direct corollary from that of the more general scheme, approximating an inhomogeneous equation, which will be constructed in Sec.5.2.1. However, our main concern now is monotonicity. The scheme given by(5.14) is not monotonic.

5.1.3 First order N scheme

Denote

$$\begin{aligned} \bar{a}_{i-\frac{1}{2},j} &= |a_{i-\frac{1}{2},j}| + |A_{i-\frac{1}{2},j+\frac{1}{2}} - A_{i-\frac{1}{2},j-\frac{1}{2}}| \\ \bar{b}_{i,j-\frac{1}{2}} &= |b_{i,j-\frac{1}{2}}| + |B_{i+\frac{1}{2},j-\frac{1}{2}} - B_{i-\frac{1}{2},j-\frac{1}{2}}|. \end{aligned} \tag{5.17}$$

Modify the previously defined upstream scheme

$$\begin{aligned} f^{u}{}_{i-\frac{1}{2},j} &= f^{c}{}_{i-\frac{1}{2},j} - \tfrac{1}{2}\bar{a}_{i-\frac{1}{2},j}(u_{i,j} - u_{i-1,j}) \\ g^{u}{}_{i,j-\frac{1}{2}} &= g^{c}{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}\bar{b}_{i,j-\frac{1}{2}}(u_{i,j} - u_{i,j-1}). \end{aligned} \tag{5.18}$$

Note that the downstream grid points may also participate with positive coefficients in the difference equation defined by (5.18). However, they will be only $O(h)$ large comparing with coefficients in the upstream grid points. Therefore, relaxation sweeps in the downstream direction will still be very efficient way to solve the difference equations. Denote

$$\begin{aligned} \alpha^{+}{}_{i-\frac{1}{2},j-\frac{1}{2}} &= \min(1, \frac{|B_{i-\frac{1}{2},j-\frac{1}{2}}|}{A^{+}{}_{i-\frac{1}{2},j-\frac{1}{2}}}) \\ \alpha^{-}{}_{i-\frac{1}{2},j+\frac{1}{2}} &= \min(1, -\frac{|B_{i-\frac{1}{2},j+\frac{1}{2}}|}{A^{-}{}_{i-\frac{1}{2},j+\frac{1}{2}}}) \end{aligned} \tag{5.19}$$

and

$$\begin{aligned} \beta^{+}{}_{i-\frac{1}{2},j-\frac{1}{2}} &= \min(1, \frac{|A_{i-\frac{1}{2},j-\frac{1}{2}}|}{B^{+}{}_{i-\frac{1}{2},j-\frac{1}{2}}}) \\ \beta^{-}{}_{i+\frac{1}{2},j-\frac{1}{2}} &= \min(1, -\frac{|A_{i+\frac{1}{2},j-\frac{1}{2}}|}{B^{-}{}_{i+\frac{1}{2},j-\frac{1}{2}}}). \end{aligned} \tag{5.20}$$

We shall call "N scheme" the first order scheme defined by

$$\begin{aligned} f^o{}_{i-\frac{1}{2},j} &= f^u{}_{i-\frac{1}{2},j} - \tfrac{1}{2}(\phi^{o+}{}_{i-\frac{1}{2},j} + \phi^{o-}{}_{i-\frac{1}{2},j}) \\ g^o{}_{i,j-\frac{1}{2}} &= g^u{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}(\gamma^{o+}{}_{i,j-\frac{1}{2}} + \gamma^{o-}{}_{i,j-\frac{1}{2}}), \end{aligned} \tag{5.21}$$

where

$$\begin{aligned} \phi^{o+}_{i-\frac{1}{2},j} &= \beta^+{}_{i-\frac{1}{2},j+\frac{1}{2}}\phi^+_{i-\frac{1}{2},j} \\ \phi^{o-}_{i-\frac{1}{2},j} &= \beta^-{}_{i-\frac{1}{2},j-\frac{1}{2}}\phi^-_{i-\frac{1}{2},j} \end{aligned} \tag{5.22}$$

and

$$\begin{aligned} \gamma^{o+}_{i,j-\frac{1}{2}} &= \alpha^+{}_{i-\frac{1}{2},j-\frac{1}{2}}\gamma^+_{i,j-\frac{1}{2}} \\ \gamma^{o-}_{i,j-\frac{1}{2}} &= \alpha^-{}_{i+\frac{1}{2},j-\frac{1}{2}}\gamma^-_{i,j-\frac{1}{2}}. \end{aligned} \tag{5.23}$$

It is easy to see that in case of a linear constant coefficient equation this scheme coincides with the N scheme defined in the previous chapter.

Theorem 5.1 *The N scheme defined by* (5.21) *is of positive type.*

Proof: When (5.21) is substituted into the balance equation, the resulting relation can be written in the following form:

$$\sum_{|k|+|l|\neq 0} C^o_{k,l}(u_{i+k,j+l} - u_{i,j}) = 0, \tag{5.24}$$

where $k, l = -1, 0, 1$. We want to show that all the coefficients $C^o_{k,l} \geq 0$ for $|k| + |l| \neq 0$. In order to show this, we have to consider all the possible situations for each gridpoint participating in the equation and to verify the non-negativity of the corresponding coefficient.

Consider the diagonal point $i-1, j-1$. The corresponding coefficient will be determined by the quantities calculated for the grid square $i-\frac{1}{2}, j-\frac{1}{2}$:

$$C^o_{-1,-1} = \frac{1}{2}(\alpha^+_{i-\frac{1}{2},j-\frac{1}{2}}A^+_{i-\frac{1}{2},j-\frac{1}{2}} + \beta^+_{i-\frac{1}{2},j-\frac{1}{2}}B^+_{i-\frac{1}{2},j-\frac{1}{2}}) \tag{5.25}$$

or

$$C^o_{-1,-1} = \min(A^+_{i-\frac{1}{2},j-\frac{1}{2}}, B^+_{i-\frac{1}{2},j-\frac{1}{2}}) \geq 0 \tag{5.26}$$

Non-negativity of other coefficients corresponding to the diagonal points can be shown in a similar way.

Consider the gridpoint $i, j-1$.

$$C^o_{0,-1} = \frac{1}{2}(b_{i,j-\frac{1}{2}} + \bar{b}_{i,j-\frac{1}{2}} - \theta^o_{i-\frac{1}{2},j-\frac{1}{2}} - \theta^o_{i+\frac{1}{2},j-\frac{1}{2}}), \tag{5.27}$$

where

$$\begin{aligned} \theta^o_{i-\frac{1}{2},j-\frac{1}{2}} &= \min(|A_{i-\frac{1}{2},j-\frac{1}{2}}|, B^+_{i-\frac{1}{2},j-\frac{1}{2}}) \\ \theta^o_{i+\frac{1}{2},j-\frac{1}{2}} &= \min(|A_{i+\frac{1}{2},j-\frac{1}{2}}|, B^+_{i+\frac{1}{2},j-\frac{1}{2}}). \end{aligned} \tag{5.28}$$

Recalling

$$\bar{b}_{i,j-\frac{1}{2}} = b_{i,j-\frac{1}{2}} + |B_{i+\frac{1}{2},j-\frac{1}{2}} - B_{i-\frac{1}{2},j-\frac{1}{2}}|, \tag{5.29}$$

it is easy to see that

$$C^o_{0,-1} \geq 0. \tag{5.30}$$

Non-negativity of other coefficients can be verified in a similar way.

5.1.4 S1 and S2 schemes

We want to add a second order correction to the N scheme, maintaining its monotonicity. We shall do it in two different ways (S1 and S2). Two difference schemes will be constructed. One of them (the S1 scheme) is slightly simpler and has the smaller truncation error, however it may admit non-physical discontinuities and is proven to be monotonic only when used with non-compressive limiters (like Roe's ψ_1 limiter). Another scheme (S2) introduces larger numerical viscosity, but rejects non-physical discontinuities and is proven to be monotonic when used with compressive limiters as well. Define

$$\begin{aligned} {\phi^{S1+}}_{i-\frac{1}{2}j} &= (1-{\beta^+}_{i-\frac{1}{2}j-\frac{1}{2}})\mathrm{sign}(A_{i-\frac{1}{2}j-\frac{1}{2}})\psi(R_{i-\frac{1}{2}j-\frac{1}{2}})B_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j-\frac{1}{2}} \\ {\phi^{S1-}}_{i-\frac{1}{2}j} &= (1-{\beta^-}_{i-\frac{1}{2}j+\frac{1}{2}})\mathrm{sign}(A_{i-\frac{1}{2}j+\frac{1}{2}})\psi(R_{i-\frac{1}{2}j+\frac{1}{2}})B_{i-\frac{1}{2}j+\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j+\frac{1}{2}} \end{aligned} \tag{5.31}$$

and

$$\begin{aligned} {\gamma^{S1+}}_{ij-\frac{1}{2}} &= (1-{\alpha^+}_{i-\frac{1}{2}j-\frac{1}{2}})\mathrm{sign}(B_{i-\frac{1}{2}j-\frac{1}{2}})\psi(Q_{i-\frac{1}{2}j-\frac{1}{2}})A_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2}j-\frac{1}{2}} \\ {\gamma^{S1-}}_{ij-\frac{1}{2}} &= (1-{\alpha^-}_{i+\frac{1}{2}j-\frac{1}{2}})\mathrm{sign}(B_{i+\frac{1}{2}j-\frac{1}{2}})\psi(Q_{i+\frac{1}{2}j-\frac{1}{2}})A_{i+\frac{1}{2}j-\frac{1}{2}}(\delta_x u)_{i+\frac{1}{2}j-\frac{1}{2}} \end{aligned} \tag{5.32}$$

where

$$R_{i-\frac{1}{2}j-\frac{1}{2}} = \frac{-A_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2}j-\frac{1}{2}}}{B_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j-\frac{1}{2}}} \tag{5.33}$$

and

$$Q_{i-\frac{1}{2}j-\frac{1}{2}} = \frac{1}{R_{i-\frac{1}{2}j-\frac{1}{2}}}. \tag{5.34}$$

Then the S1 scheme can be defined by

$$\begin{aligned} {f^{S1}}_{i-\frac{1}{2}j} &= {f^o}_{i-\frac{1}{2}j} - \tfrac{1}{2}({\phi^{S1+}}_{i-\frac{1}{2}j} + {\phi^{S1-}}_{i-\frac{1}{2}j}) \\ {g^{S1}}_{ij-\frac{1}{2}} &= {g^o}_{ij-\frac{1}{2}} - \tfrac{1}{2}({\gamma^{S1+}}_{ij-\frac{1}{2}} + {\gamma^{S1-}}_{ij-\frac{1}{2}}) \end{aligned} \tag{5.35}$$

We can formulate the following monotonicity result

Theorem 5.2 *The S1 scheme defined by* (5.35) *is monotonic if*

$$0 \leq \frac{\psi(R)}{R} \leq 1, \ \psi(R) \leq 1. \tag{5.36}$$

Proof: Using the following identities

$$\begin{aligned}\psi(R_{i-\frac{1}{2},j-\frac{1}{2}})B_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} &\equiv -\frac{\psi(R_{i-\frac{1}{2},j-\frac{1}{2}})}{R_{i-\frac{1}{2},j-\frac{1}{2}}}A_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2},j-\frac{1}{2}}\\ \psi(Q_{i-\frac{1}{2},j-\frac{1}{2}})A_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2},j-\frac{1}{2}} &\equiv -\frac{\psi(Q_{i-\frac{1}{2},j-\frac{1}{2}})}{Q_{i-\frac{1}{2},j-\frac{1}{2}}}B_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}}\end{aligned} \tag{5.37}$$

it can be verified that it is always possible to rewrite the S1 scheme in the form

$$\sum_{|k|+|l|\neq 0} C^{S1}_{k,l}(u_{i+k,j+l}-u_{i,j}) = 0, \tag{5.38}$$

where $k,l=-1,0,1$ and, provided (5.36) holds $C^{S1}_{k,l}\geq 0$ for $|k|+|l|\neq 0$.

The complete proof is straightforward. We shall only illustrate the importance of the restriction (5.36) on two typical situations, considering grid point $(i,j-1)$.

1. Suppose

$$A_{i-\frac{1}{2},j-\frac{1}{2}} = B_{i-\frac{1}{2},j-\frac{1}{2}} > 0 \tag{5.39}$$

This means that

$$\begin{aligned}(\delta_x u)_{i-\frac{1}{2},j-\frac{1}{2}} &= u_{i,j-1}-u_{i-1,j-1}\\ (\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} &= u_{i-1,j}-u_{i-1,j-1}.\end{aligned} \tag{5.40}$$

Suppose also

$$B_{i+\frac{1}{2},j-\frac{1}{2}} = B_{i-\frac{1}{2},j-\frac{1}{2}} \tag{5.41}$$

and that $A_{i+\frac{1}{2},j-\frac{1}{2}}$ is a small positive number. Then

$$\begin{aligned}(\delta_x u)_{i+\frac{1}{2},j-\frac{1}{2}} &= u_{i+1,j-1}-u_{i,j-1}\\ (\delta_y u)_{i+\frac{1}{2},j-\frac{1}{2}} &= u_{i,j}-u_{i,j-1}.\end{aligned} \tag{5.42}$$

This yields

$$C^{S1}_{0,-1} = C^{o}_{0,-1} - (1-\beta^{+}_{i+\frac{1}{2},j-\frac{1}{2}})\psi(R_{i+\frac{1}{2},j-\frac{1}{2}})B_{i+\frac{1}{2},j-\frac{1}{2}} \tag{5.43}$$

or

$$\begin{aligned}C^{S1}_{0,-1} = \frac{1}{2}(b_{i,j-\frac{1}{2}} &+ b_{i,j-\frac{1}{2}} - B_{i-\frac{1}{2},j-\frac{1}{2}} - A_{i+\frac{1}{2},j-\frac{1}{2}}\\ &-(B_{i+\frac{1}{2},j-\frac{1}{2}} - A_{i+\frac{1}{2},j-\frac{1}{2}})\psi(R_{i+\frac{1}{2},j-\frac{1}{2}}))\end{aligned} \tag{5.44}$$

The last expression will be non-negative if (5.36) holds, but may attain negative values if $\psi(R)>1$.

2. Suppose that $B_{i-\frac{1}{2},j-\frac{1}{2}}>0$ and that $A_{i-\frac{1}{2},j-\frac{1}{2}}<0$ has a small absolute value. This means that

$$\begin{aligned}(\delta_x u)_{i-\frac{1}{2},j-\frac{1}{2}} &= u_{i,j-1}-u_{i-1,j-1}\\ (\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} &= u_{i,j}-u_{i,j-1}.\end{aligned} \tag{5.45}$$

Suppose also

$$B_{i+\frac{1}{2},j-\frac{1}{2}} = B_{i-\frac{1}{2},j-\frac{1}{2}} \tag{5.46}$$

and $A_{i+\frac{1}{2},j-\frac{1}{2}}$ is a small positive number. Then

$$\begin{aligned} (\delta_x u)_{i+\frac{1}{2},j-\frac{1}{2}} &= u_{i+1,j-1} - u_{i,j-1} \\ (\delta_y u)_{i+\frac{1}{2},j-\frac{1}{2}} &= u_{i,j} - u_{i,j-1}. \end{aligned} \tag{5.47}$$

Note that this case corresponds to a rarefaction wave. Here we obtain

$$\begin{aligned} C^{S1}_{0,-1} = \frac{1}{2}(b_{i,j-\frac{1}{2}} + \bar{b}_{i,j-\frac{1}{2}} - (1 - \beta^+_{i-\frac{1}{2},j-\frac{1}{2}})\psi(R_{i-\frac{1}{2},j-\frac{1}{2}})B_{i-\frac{1}{2},j-\frac{1}{2}} \\ -(1 - \beta^+_{i+\frac{1}{2},j-\frac{1}{2}})\psi(R_{i+\frac{1}{2},j-\frac{1}{2}})B_{i+\frac{1}{2},j-\frac{1}{2}} \end{aligned} \tag{5.48}$$

Since $\beta^+_{i-\frac{1}{2},j-\frac{1}{2}}$ and $\beta^+_{i+\frac{1}{2},j-\frac{1}{2}}$ are small, the last expression may attain negative values if $\psi(R) > 1$. However, it will be non-negative, if (5.36) holds. If we rewrite the S1 scheme corrections using the identities (5.37), it will create negative coefficients at diagonal grid points $(i-1, j-1)$ and $(i+1, j-1)$.

Remark 5.1 The inequality (5.36) means that no artificial compression can be added to the schemes.

Remark 5.2 The S1 scheme (as well as the N scheme) will perfectly resolve contact discontinuities and shock waves which align either with grid lines or with grid diagonals. However, non-physical discontinuities (corresponding to rarefaction waves) can also be admitted in these cases. This is because of the vanishing cross-stream truncation error.

We shall now modify the S1 scheme in order to enable it to reject non-physical discontinuities and to allow the addition of artificial compression when needed. Denote

$$\begin{aligned} (\varrho_x A)_{i-\frac{1}{2},j-\frac{1}{2}} &= |a_{i,j-\nu} - a_{i-1,j-\nu}| \\ (\varrho_y B)_{i-\frac{1}{2},j-\frac{1}{2}} &= |b_{i-\mu,j} - b_{i-\mu,j-1}|, \end{aligned} \tag{5.49}$$

where μ and ν are the same as in definitions of $A_{i-\frac{1}{2},j-\frac{1}{2}}$ and $B_{i-\frac{1}{2},j-\frac{1}{2}}$ (5.9). Denote also

$$(\delta^2_{xy} u)_{i-\frac{1}{2},j-\frac{1}{2}} = u_{i,j} - u_{i-1,j} - u_{i,j-1} + u_{i-1,j-1} \tag{5.50}$$

$$\begin{aligned} \phi^{e+}{}_{i-\frac{1}{2},j} &= 2(\varrho_y B)_{i-\frac{1}{2},j-\frac{1}{2}}(u_{i,j} - u_{i-1,j}) - (\varrho_x A)_{i-\frac{1}{2},j-\frac{1}{2}}(\delta^2_{xy} u)_{i-\frac{1}{2},j-\frac{1}{2}} \\ \phi^{e-}{}_{i-\frac{1}{2},j} &= 2(\varrho_y B)_{i-\frac{1}{2},j-\frac{1}{2}}(u_{i,j} - u_{i-1,j}) + (\varrho_x A)_{i-\frac{1}{2},j+\frac{1}{2}}(\delta^2_{xy} u)_{i-\frac{1}{2},j+\frac{1}{2}} \end{aligned} \tag{5.51}$$

$$\begin{aligned} \gamma^{e+}{}_{i,j-\frac{1}{2}} &= 2(\varrho_x A)_{i-\frac{1}{2},j-\frac{1}{2}}(u_{i,j} - u_{i,j-1}) - (\varrho_y B)_{i-\frac{1}{2},j-\frac{1}{2}}(\delta^2_{xy} u)_{i-\frac{1}{2},j-\frac{1}{2}} \\ \gamma^{e-}{}_{i,j-\frac{1}{2}} &= 2(\varrho_x A)_{i+\frac{1}{2},j-\frac{1}{2}}(u_{i,j} - u_{i,j-1}) + (\varrho_y B)_{i+\frac{1}{2},j-\frac{1}{2}}(\delta^2_{xy} u)_{i+\frac{1}{2},j-\frac{1}{2}} \end{aligned} \tag{5.52}$$

Define

$$\begin{aligned} f^{S2}{}_{i-\frac{1}{2},j} &= f^{S1}{}_{i-\frac{1}{2},j} - (\phi^{e+}{}_{i-\frac{1}{2},j} + \phi^{e-}{}_{i-\frac{1}{2},j}) \\ g^{S2}{}_{i,j-\frac{1}{2}} &= g^{S1}{}_{i,j-\frac{1}{2}} - (\gamma^{e+}{}_{i,j-\frac{1}{2}} + \gamma^{e-}{}_{i,j-\frac{1}{2}}) \end{aligned} \tag{5.53}$$

Theorem 5.3 *The S2 scheme is monotonic if*

$$0 \leq \frac{\psi(R)}{R} \leq 2, \ \psi(R) \leq 2. \tag{5.54}$$

Proof: Once again we claim that it is possible to rewrite the S scheme using the identities (5.37), where neccessary, in the following form

$$\sum_{|k|+|l|\neq 0} C^{S2}_{k,l}(u_{i+k,j+l} - u_{i,j}) = 0, \tag{5.55}$$

where $k, l = -1, 0, 1$ and $C^{S2}_{k,l} \geq 0$ for $|k| + |l| \neq 0$.

We shall decsribe in details only the same two possible situations, as in the previous theorem, considering a grid point $(i, j-1)$:

1. Suppose

$$A_{i-\frac{1}{2},j-\frac{1}{2}} = B_{i-\frac{1}{2},j-\frac{1}{2}} > 0, \tag{5.56}$$

$$B_{i+\frac{1}{2},j-\frac{1}{2}} = B_{i-\frac{1}{2},j-\frac{1}{2}} \tag{5.57}$$

and $A_{i+\frac{1}{2},j-\frac{1}{2}}$ a small positive number.

Then

$$C^{S2}_{0,-1} = C^{S1}_{0,-1} + ((\varrho_x A)_{i-\frac{1}{2},j-\frac{1}{2}} + (\varrho_x A)_{i+\frac{1}{2},j-\frac{1}{2}}) \tag{5.58}$$

Recalling the expression for $C_{0,-1}$ and taking into account the following

$$((\varrho_x A)_{i-\frac{1}{2},j-\frac{1}{2}} + (\varrho_x A)_{i+\frac{1}{2},j-\frac{1}{2}}) \geq |A_{i-\frac{1}{2},j-\frac{1}{2}} - A_{i+\frac{1}{2},j-\frac{1}{2}}|, \tag{5.59}$$

we can conclude that $C^{S2}_{0,-1}$ is nonnegative if (5.54) holds.

2. Suppose

$$B_{i+\frac{1}{2},j-\frac{1}{2}} = B_{i-\frac{1}{2},j-\frac{1}{2}} \tag{5.60}$$

and $A_{i+\frac{1}{2},j-\frac{1}{2}}$ is a small positive number.

Suppose also thatl $B_{i-\frac{1}{2},j-\frac{1}{2}} > 0$ and $A_{i-\frac{1}{2},j-\frac{1}{2}} < 0$ has a small absolute valuer. Note that, since $A_{i-\frac{1}{2},j-\frac{1}{2}}$ and $A_{i+\frac{1}{2},j-\frac{1}{2}}$ have opposite signs, then either

$$(\varrho_x A)_{i-\frac{1}{2},j-\frac{1}{2}} \geq |A_{i-\frac{1}{2},j-\frac{1}{2}}| \tag{5.61}$$

or

$$(\varrho_x A)_{i+\frac{1}{2},j-\frac{1}{2}} \geq A_{i+\frac{1}{2},j-\frac{1}{2}}. \tag{5.62}$$

Suppose for simplicity, that the first is correct. Then the S1 scheme correction corresponding to $A_{i-\frac{1}{2},j-\frac{1}{2}}$ and $B_{i-\frac{1}{2},j-\frac{1}{2}}$ can be substituted by using (5.37) and we shall obtain

$$C^{S2}_{-1,-1} = (\varrho_y B)_{i-\frac{1}{2},j-\frac{1}{2}} + (\varrho_x A)_{i-\frac{1}{2},j-\frac{1}{2}} - \frac{1}{2}\frac{\psi(R_{i-\frac{1}{2},j-\frac{1}{2}})}{R_{i-\frac{1}{2},j-\frac{1}{2}}} A_{i-\frac{1}{2},j-\frac{1}{2}} \tag{5.63}$$

and

$$\begin{aligned} C^{S2}_{0,-1} &= \frac{1}{2}(b_{i,j-\frac{1}{2}} + \bar{b}_{i,j-\frac{1}{2}} \\ &+(1 - \beta^{+}_{i-\frac{1}{2},j-\frac{1}{2}})\frac{\psi(R_{i-\frac{1}{2},j-\frac{1}{2}})}{R_{i-\frac{1}{2},j-\frac{1}{2}}}A_{i-\frac{1}{2},j-\frac{1}{2}} \\ &-(1 - \beta^{+}_{i+\frac{1}{2},j-\frac{1}{2}})\psi(R_{i+\frac{1}{2},j-\frac{1}{2}})B_{i+\frac{1}{2},j-\frac{1}{2}}) \end{aligned}$$

It is easy to see that both $C^{S2}_{-1,-1}$ and $C^{S2}_{0,-1}$ will be nonnegative if (5.54) holds.

Remark 5.3 The S2 scheme which was created from the S1 scheme by adding the e-correction allows us to use limiters which may introduce artificial compression.

Remark 5.4 Another important property of the e-correction is, that it will introduce additional cross-stream viscosity in case of a rarefaction wave (diverging characteristic field). Therefore, no non-physical discontinuities will be admitted.

5.2 Inhomogeneous equation

Consider an inhomogeneous conservation law

$$(f(u))_x + (g(u))_y = s, \tag{5.64}$$

where $s = s(x, y)$.

5.2.1 2D scheme

The same central and upstream approximations can be used in this case and they will be second and first order accurate respectively. However, the genuinely 2D scheme has to be modified, in order to maintain the second order accuracy.

Denote

$$\begin{aligned} s_{i-\frac{1}{2}j} &= \tfrac{1}{2}(s_{ij} + s_{i-1,j}) \\ s_{ij-\frac{1}{2}} &= \tfrac{1}{2}(s_{ij} + s_{ij-1}) \end{aligned} \tag{5.65}$$

Define

$$s^{+}_{i-\frac{1}{2}j} = \frac{s_{i-\frac{1}{2}j}B^{+}_{i-\frac{1}{2}j-\frac{1}{2}}}{B^{+}_{i-\frac{1}{2}j-\frac{1}{2}} - B^{-}_{i-\frac{1}{2}j+\frac{1}{2}}} \tag{5.66}$$

$$s^{-}_{i-\frac{1}{2}j} = \frac{-s_{i-\frac{1}{2}j}B^{-}_{i-\frac{1}{2}j+\frac{1}{2}}}{B^{+}_{i-\frac{1}{2}j-\frac{1}{2}} - B^{-}_{i-\frac{1}{2}j+\frac{1}{2}}} \tag{5.67}$$

$$s^+_{i,j-\frac{1}{2}} = \frac{s_{i-\frac{1}{2},j}A^+_{i-\frac{1}{2},j-\frac{1}{2}}}{A^+_{i-\frac{1}{2},j-\frac{1}{2}} - A^-_{i+\frac{1}{2},j-\frac{1}{2}}} \tag{5.68}$$

$$s^-_{i,j-\frac{1}{2}} = \frac{-s_{i-\frac{1}{2},j}A^-_{i+\frac{1}{2},j-\frac{1}{2}}}{A^+_{i-\frac{1}{2},j-\frac{1}{2}} - A^-_{i+\frac{1}{2},j-\frac{1}{2}}} \tag{5.69}$$

Again

$$\begin{aligned} f^{2D}{}_{i-\frac{1}{2},j} &= f^u{}_{i-\frac{1}{2},j} - \tfrac{1}{2}(\phi^+_{i-\frac{1}{2},j} + \phi^-_{i-\frac{1}{2},j}) \\ g^{2D}{}_{i,j-\frac{1}{2}} &= g^u{}_{i,j-\frac{1}{2}} - \tfrac{1}{2}(\gamma^+_{i,j-\frac{1}{2}} + \gamma^-_{i,j-\frac{1}{2}}), \end{aligned} \tag{5.70}$$

but

$$\begin{aligned} \phi^+_{i-\frac{1}{2},j} &= \mathrm{sign}(A_{i-\frac{1}{2},j-\frac{1}{2}})(B^+{}_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} + hs^+_{i-\frac{1}{2},j}) \\ \phi^-_{i-\frac{1}{2},j} &= \mathrm{sign}(A_{i-\frac{1}{2},j+\frac{1}{2}})(B^-{}_{i-\frac{1}{2},j+\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j+\frac{1}{2}} + hs^-_{i-\frac{1}{2},j}) \end{aligned} \tag{5.71}$$

$$\begin{aligned} \gamma^+_{i,j-\frac{1}{2}} &= \mathrm{sign}(B_{i-\frac{1}{2},j-\frac{1}{2}})(A^+{}_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2},j-\frac{1}{2}} + hs^+_{i,j-\frac{1}{2}}) \\ \gamma^-_{i,j-\frac{1}{2}} &= \mathrm{sign}(B_{i+\frac{1}{2},j-\frac{1}{2}})(A^-{}_{i+\frac{1}{2},j-\frac{1}{2}}(\delta_x u)_{i+\frac{1}{2},j-\frac{1}{2}} + hs^-_{i,j-\frac{1}{2}}) \end{aligned} \tag{5.72}$$

Theorem 5.4 *The 2D scheme defined by* (5.70) *with* (5.71) *and* (5.72) *is second order accurate.*

Proof: Assume that

$$\begin{aligned} \mathrm{sign}(A_{i-\frac{1}{2},j-\frac{1}{2}}) = \mathrm{sign}(A_{i-\frac{1}{2},j+\frac{1}{2}}) &= \mathrm{sign}(A_{i+\frac{1}{2},j-\frac{1}{2}}) = \mathrm{sign}(A_{i+\frac{1}{2},j+\frac{1}{2}}) \\ \mathrm{sign}(B_{i-\frac{1}{2},j-\frac{1}{2}}) = \mathrm{sign}(B_{i-\frac{1}{2},j+\frac{1}{2}}) &= \mathrm{sign}(B_{i+\frac{1}{2},j-\frac{1}{2}}) = \mathrm{sign}(B_{i+\frac{1}{2},j+\frac{1}{2}}). \end{aligned} \tag{5.73}$$

Then

$$\begin{aligned} (\phi^+_{i+\frac{1}{2},j} + \phi^-_{i+\frac{1}{2},j}) &- (\phi^+_{i-\frac{1}{2},j} + \phi^-_{i-\frac{1}{2},j}) \\ = h^2\mathrm{sign}(a(u))(b(u)u_y &- s)_x + O(h^3). \end{aligned} \tag{5.74}$$

Assuming that

$$\mathrm{sign}(A_{i-\frac{1}{2},j+\frac{1}{2}} - A_{i-\frac{1}{2},j+\frac{1}{2}}) = \mathrm{sign}(A_{i+\frac{1}{2},j+\frac{1}{2}} - A_{i+\frac{1}{2},j-\frac{1}{2}}), \tag{5.75}$$

we also obtain

$$\begin{aligned} \frac{1}{2}(|\bar{a}_{i+\frac{1}{2},j}|(u_{i+1,j} - u_{i,j}) &- \frac{1}{2}(|\bar{a}_{i-\frac{1}{2},j}|(u_{i,j} - u_{i-1,j}) \\ = \frac{1}{2}(h^2\mathrm{sign}(a(u))(a(u)u_x)_x &+ h^3\mathrm{sign}(a_y)(a_y u_x)_x) + O(h^3) \end{aligned} \tag{5.76}$$

Since the central scheme is second order accurate

$$\begin{aligned} f^{2D}_{i+\frac{1}{2}j} - f^{2D}_{i-\frac{1}{2}j} &= (f(u))_x + O(h^3) \\ +\frac{1}{2}(|\bar{a}_{i+\frac{1}{2}j}|(u_{i+1j} - u_{ij}) &- (\phi^+_{i+\frac{1}{2}j} + \phi^-_{i+\frac{1}{2}j})) \\ -\frac{1}{2}(|\bar{a}_{i-\frac{1}{2}j}|(u_{ij} - u_{i-1j}) &- (\phi^+_{i-\frac{1}{2}j} + \phi^-_{i-\frac{1}{2}j})). \end{aligned} \tag{5.77}$$

Taking into account the previous three relations, we conclude

$$\begin{aligned} & f^{2D}_{i+\frac{1}{2}j} - f^{2D}_{i-\frac{1}{2}j} \\ = (f(u))_x + h^2 \mathrm{sign}(a(u))(a(u)u_x &+ b(u)u_y - s)_x + O(h^3) = O(h^3) \end{aligned} \tag{5.78}$$

Similarly

$$\begin{aligned} & g^{2D}_{ij+\frac{1}{2}} - g^{2D}_{ij-\frac{1}{2}} \\ = (g(u))_x + h^2 \mathrm{sign}(b(u))(a(u)u_x &+ b(u)u_y - s)_y + O(h^3) = O(h^3) \end{aligned} \tag{5.79}$$

If one of the assumptions (5.73) or (5.75) does not hold, the difference equation in this computational cell will approximate the differential one only up to $O(h)$. However, this can happen in computational cells which cover only $O(h)$ part of the domain. Therefore, the scheme will be still second order accurate.

5.2.2 N scheme

If we are interested in obtaining first order accuracy only, we can use the same N scheme, as for the homogeneous case. However, if the N scheme is an intermediate step towards constructing higher order schemes, it also has to be modified.

Once again

$$\begin{aligned} f^o{}_{i-\frac{1}{2}j} &= f^u{}_{i-\frac{1}{2}j} - \tfrac{1}{2}(\phi^{o+}{}_{i-\frac{1}{2}j} + \phi^{o-}{}_{i-\frac{1}{2}j}) \\ g^o{}_{ij-\frac{1}{2}} &= g^u{}_{ij-\frac{1}{2}} - \tfrac{1}{2}(\gamma^{o+}{}_{ij-\frac{1}{2}} + \gamma^{o-}{}_{ij-\frac{1}{2}}) \end{aligned} \tag{5.80}$$

where

$$\begin{aligned} \phi^{o+}_{i-\frac{1}{2}j} &= \beta^+{}_{i-\frac{1}{2}j+\frac{1}{2}}\phi^+_{i-\frac{1}{2}j} \\ \phi^{o-}_{i-\frac{1}{2}j} &= \beta^-{}_{i-\frac{1}{2}j-\frac{1}{2}}\phi^-_{i-\frac{1}{2}j} \end{aligned} \tag{5.81}$$

$$\begin{aligned} \gamma^{o+}_{ij-\frac{1}{2}} &= \alpha^+{}_{i-\frac{1}{2}j-\frac{1}{2}}\gamma^+_{ij-\frac{1}{2}} \\ \gamma^{o-}_{ij-\frac{1}{2}} &= \alpha^-{}_{i+\frac{1}{2}j-\frac{1}{2}}\gamma^-_{ij-\frac{1}{2}} \end{aligned} \tag{5.82}$$

5.2.3 S scheme

Define

$$R^+_{i-\frac{1}{2}j} = \frac{-A_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2}j-\frac{1}{2}}}{B_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j-\frac{1}{2}} + hs_{i-\frac{1}{2}j}} \tag{5.83}$$

$$R^-_{i-\frac{1}{2}j} = \frac{-A_{i-\frac{1}{2}j+\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2}j+\frac{1}{2}}}{B_{i-\frac{1}{2}j+\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j+\frac{1}{2}} + hs_{i-\frac{1}{2}j}} \tag{5.84}$$

$$Q^+_{ij-\frac{1}{2}} = \frac{-B_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j-\frac{1}{2}}}{A_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2}j-\frac{1}{2}} + hs_{ij-\frac{1}{2}}} \tag{5.85}$$

$$Q^-_{ij-\frac{1}{2}} = \frac{-B_{i+\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i+\frac{1}{2}j-\frac{1}{2}}}{A_{i+\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i+\frac{1}{2}j-\frac{1}{2}} + hs_{ij-\frac{1}{2}}} \tag{5.86}$$

Then, the S1 scheme can be given by

$$\begin{aligned} f^{S1}{}_{i-\frac{1}{2}j} &= f^\circ{}_{i-\frac{1}{2}j} - \tfrac{1}{2}(\phi^{S1+}{}_{i-\frac{1}{2}j} + \phi^{S1-}{}_{i-\frac{1}{2}j}) \\ g^{S1}{}_{ij-\frac{1}{2}} &= g^\circ{}_{ij-\frac{1}{2}} - \tfrac{1}{2}(\gamma^{S1+}{}_{ij-\frac{1}{2}} + \gamma^{S1-}{}_{ij-\frac{1}{2}}) \end{aligned} \tag{5.87}$$

with

$$\begin{aligned} &\phi^{S1+}{}_{i-\frac{1}{2}j} = \\ &\quad (1- \beta^+{}_{i-\frac{1}{2}j-\frac{1}{2}})\psi(R^+_{i-\frac{1}{2}j})\mathrm{sign}(A_{i-\frac{1}{2}j-\frac{1}{2}})(B_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j-\frac{1}{2}} + s_{i-\frac{1}{2}j}) \\ &\phi^{S1-}{}_{i-\frac{1}{2}j} = \\ &\quad (1- \beta^-{}_{i-\frac{1}{2}j+\frac{1}{2}})\psi(R^-_{i-\frac{1}{2}j})\mathrm{sign}(A_{i-\frac{1}{2}j+\frac{1}{2}})(B_{i-\frac{1}{2}j+\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j+\frac{1}{2}} + s_{i-\frac{1}{2}j}) \end{aligned} \tag{5.88}$$

$$\begin{aligned} &\gamma^{S1+}{}_{ij-\frac{1}{2}} = \\ &\quad (1- \alpha^+{}_{i-\frac{1}{2}j-\frac{1}{2}})\psi(Q^+_{ij-\frac{1}{2}})\mathrm{sign}(B_{i-\frac{1}{2}j-\frac{1}{2}})(A_{i-\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2}j-\frac{1}{2}} + s_{ij-\frac{1}{2}}) \\ &\gamma^{S1-}{}_{ij-\frac{1}{2}} = \\ &\quad (1- \alpha^-{}_{i+\frac{1}{2}j-\frac{1}{2}})\psi(Q^-_{ij-\frac{1}{2}})\mathrm{sign}(B_{i+\frac{1}{2}j-\frac{1}{2}})(A_{i+\frac{1}{2}j-\frac{1}{2}}(\delta_y u)_{i+\frac{1}{2}j-\frac{1}{2}} + s_{ij-\frac{1}{2}}). \end{aligned} \tag{5.89}$$

Theorem 5.5 *If $\psi = \psi(R) \in C^2$ and*

$$\psi(1) = 1, \tag{5.90}$$

then the S1 scheme is second order accurate.

Proof: In addition to (5.75) and (5.73), assume also that

$$B_{i-\frac{1}{2}j-\frac{1}{2}}, B_{i-\frac{1}{2}j+\frac{1}{2}} > 0. \tag{5.91}$$

This means that

$$\begin{aligned} B^+_{i-\frac{1}{2},j-\frac{1}{2}} &= B_{i-\frac{1}{2},j-\frac{1}{2}} > 0 \\ B^-_{i-\frac{1}{2},j+\frac{1}{2}} &= 0, \end{aligned} \tag{5.92}$$

$$\begin{aligned} \phi^-_{i-\frac{1}{2},j} &= 0 \\ \phi^{S-}_{i-\frac{1}{2},j} &= 0 \end{aligned} \tag{5.93}$$

and

$$\begin{aligned} s^-_{i-\frac{1}{2},j} &= 0 \\ s^+_{i-\frac{1}{2},j} &= s_{i-\frac{1}{2},j}. \end{aligned} \tag{5.94}$$

Then

$$\begin{aligned} f^{2D}_{i-\frac{1}{2},j} &- f^{S1}_{i-\frac{1}{2},j} = \\ -\frac{1}{2}(1 &- \psi(R^+_{i-\frac{1}{2},j})\mathrm{sign}(A_{i-\frac{1}{2},j-\frac{1}{2}})(1-\beta^+_{i-\frac{1}{2},j}) \\ (B_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} &+ hs_{i-\frac{1}{2},j}) \end{aligned} \tag{5.95}$$

Taking into account

$$1-\psi(R) = \psi'_R(1)(R-1) + O(h^2), \tag{5.96}$$

we have

$$\begin{aligned} f^{2D}_{i-\frac{1}{2},j} &- f^{S1}_{i-\frac{1}{2},j} = \\ -\frac{1}{2}\psi'_R(1)(R^+_{i-\frac{1}{2},j-\frac{1}{2}} &- 1)\mathrm{sign}(A_{i-\frac{1}{2},j-\frac{1}{2}})(1-\beta^+_{i-\frac{1}{2},j}) \\ (B_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} &+ hs_{i-\frac{1}{2},j}) + O(h^3) \end{aligned} \tag{5.97}$$

Recalling the definition of $R^+_{i-\frac{1}{2},j-\frac{1}{2}}$, we can rewrite (5.97) as

$$\begin{aligned} f^{2D}_{i-\frac{1}{2},j} &- f^{S1}_{i-\frac{1}{2},j} = \\ -\frac{1}{2}\psi'_R(1)(-A_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_x u)_{i-\frac{1}{2},j-\frac{1}{2}} &- B_{i-\frac{1}{2},j-\frac{1}{2}}(\delta_y u)_{i-\frac{1}{2},j-\frac{1}{2}} - hs_{i-\frac{1}{2},j}) \\ \mathrm{sign}(A_{i-\frac{1}{2},j-\frac{1}{2}})(1 &- \beta^+_{i-\frac{1}{2},j}) + O(h^3). \end{aligned}$$

Then

$$f^{S1}_{i+\frac{1}{2},j} - f^{S1}_{i-\frac{1}{2},j} = f^{2D}_{i+\frac{1}{2},j} - f^{2D}_{i-\frac{1}{2},j} + O(h^3). \tag{5.98}$$

If one of the assumptions (5.73), (5.75) or (5.91) does not hold, the difference equation at that computational cell will approximate the differential one only upto $O(h)$. However, this can happen in computational cells which cover only $O(h)$ part of the domain. Therefore, the scheme will be still second order accurate.

Remark 5.5 The S2 scheme created by adding the e-correction to the S1 scheme will also be second order accurate. This is because the e-correction is $O(h^2)$.

Chapter 6

Numerical experiments

All the numerical experiments reported here deal with numerical solutions of the differential equation

$$-(0+)\Delta u + (f(u))_x + (g(u))_y = s, \tag{6.1}$$

where 0+ means an infinitesimally small positive number.

We choose for our domain the rectangle :

$$\Omega = \{(x,y) : 0 \le x \le 3, 0 \le y \le 2\} \tag{6.2}$$

and we set Dirichlet boundary conditions around its boundary.

We used five qrids (levels). The meshsize of grid k is $h_k = 2^{1-k}$, hence it has $(3 \times 2^{k-1} - 1) \times (2^k - 1)$ interior grid points and 5×2^k boundary points $(1 \le k \le 5)$.

Note that in all the cases reported below limit solutions can be obtained just by few downstream relaxation sweeps on the finest grid. This is because all the difference schemes used are upstream (or "almost" upstream). However, Eq.(6.1) is just a model problem for more complicated systems which cannot be solved this way. In case of subsonic flow for instance the equations contain an elliptic component as well as hyperbolic. In case of supersonic flow there exist several families of characteristics. Therefore, the main purpose of the experiments reported here is to demonstrate that the developed discretization schemes provide the possibility to obtain second order accurate solution (in smooth regions and also in terms of discontinuity location) by means of a certain multigrid algoritm, employing a *direction-free* relaxation.

The algorithm used is of type $FMG(N, N_M, C, M)$ (see Chap.3.), where $M = 5, N = 1,2$, $N_M = 2,6$, and $C = W(2,1)$. The Full-Weighted residual transfer is used. In case the N, S, S1 or S2 schemes are used, the Red-Black relaxation without storing of intermediate values is not direction-free anymore. Therefore, we use "4-colour" ordering. The usual bilinear correction interpolation and bicubic FMG interpolation is employed.

The precise formulas for numerical flux derivatives were used in all the experiments reported here. However, there will not be any significant difference in the performance of the algorithms if the N scheme numerical flux derivative formulas will be used for the S,S1 and S2 schemes. This has been observed in a comparison of the solution error and residual behaviour in both cases.

We shall compare the solutions obtained by different discretizations and discuss the choice of the difference scheme for a particular problem.

6.1 Linear equation

We shall first examine the performance of the algorithms and the quality of obtained numerical solutions in case of the linear equation

$$-(0+)\Delta u + au_x + bu_y = s. \tag{6.3}$$

6.1.1 Smooth solution

First we study how effective different algorithms are in case of smooth solutions, avoiding any influence of discontinuities. For this purpose we provide equation (6.3) with such boundary conditions that there will be no boundary layers. Since all the difference schemes we experiment with here are upstream, we do not expect the appearance of even a numerical boundary layer.

Homogeneous case

Consider the following version of equation (6.3)

$$-(0+)\Delta u + .5u_x + u_y = 0, \tag{6.4}$$

with boundary conditions given by

$$u = \sin(y - 2x). \tag{6.5}$$

It is easy to see that (6.5) is also the exact solution of (6.4). This solution is constant along the characteristics and varies in other directions. We want first to compare performance of the upstream, N, 2D and S schemes on this model problem. The S scheme is employed here in three versions: with Van Leer's limiter and two Roe's limiters. The experiments with this model problem are presented in Table 6.1.

Each column of Table 6.1 starting from the second presents the history of the L_1 norm of the solution error (the difference between the numerical and the differential solution). $2FMG$ algorithm is employed (i.e., $N = 2$). The first column indicates the stage of the multigrid algorithm the error is displayed at: the first number stands for the currently finest level and the number in parentheses says how many multigrid cycles have already been performed on this level (zero means that the error is displayed just after the bicubic interpolation to this level). Columns 1,2 correspond to the cases where the upstream and N schemes are used respectively. Both these schemes demonstrate first order convergence. However, the solution error in case of the N scheme is almost three times smaller. The $1FMG$ algorithm will produce results of the same quality in both these cases (This can be seen , e.g., by comparing the results of row 5(1) with those of 5(2) and 5(6); the latter practically show the discretization error). Column 3 presents the experiments with the 2D scheme and Columns 4,5 and 6 correspond to the experiments

Difference scheme	Upstr	N	2D	S	S	S
Limiter	–	–	–	ψ_{VL}	ψ_1	ψ_2
2(5)	.241	.119	.0429	.0497	.0588	.0471
3(2)	.154	.0632	.00971	.0145	.0232	.0191
4(2)	.0870	.0326	.00251	.00398	.00688	.00638
5(0)	.0846	.0317	.00243	.00376	.00656	.00597
5(1)	.0475	.0161	.00054	.00150	.00266	.00275
5(2)	.0467	.0165	.00065	.00113	.00213	.00235
⋮						
5(6)	.0468	.0166	.00060	.00097	.00194	.00219
	1	2	3	4	5	6

Table 6.1: Linear homogeneous problem; *slowly varying solution.*

with Van Leer's limiter and Roe's ψ_1 and ψ_2 limiters respectively. The second order convergence can be observed in all these cases.

Table 6.2 presents the experiments with the same equation (6.4) but with different, more oscillatory boundary conditions:

$$u = \sin(5(y - 2x)) \tag{6.6}$$

and has the same structure as Table 6.1.

Difference scheme	Upstr	N	2D	S	S	S
Limiter	–	–	–	ψ_{VL}	ψ_1	ψ_2
2(5)	.721	.759	.972	.779	.771	.788
3(2)	.644	.577	.860	.557	.561	.553
4(2)	.584	.434	.382	.267	.317	.202
5(0)	.569	.429	.358	.274	.322	.208
5(1)	.515	.327	.131	.0994	.155	.0806
5(2)	.496	.298	.123	.0728	.121	.0705
⋮						
5(6)	.490	.294	.0835	.0675	.113	.0542
	1	2	3	4	5	6

Table 6.2: Linear homogeneous problem; oscillatory solution.

The upstream and N schemes seem to give a very poor approximation to such a solution, however the 2D and S schemes with all the limiters start to demonstrate a

second order convergence on the finer levels. However, about 3 cycles are needed on the finest level to achieve a second order accurate solution. Performing more than 2 cycles on each coarser level does not change the situation, because the coarse grid solution provides a poor approximation for the finer grid one (for detailed analysis of this situation see Sec.2.3 in [4]).

Inhomogeneous case

Consider the following inhomogeneous equation

$$-(0+)\Delta u + .5u_x + u_y = s, \tag{6.7}$$

with boundary conditions

$$u = \sin(x+y). \tag{6.8}$$

Choose the right-hand-side to be

$$s = 1.5\cos(x+y). \tag{6.9}$$

Then (6.8) will be the exact solution of (6.7).

Table 6.3 has also the same structure as Table 6.1.

Difference scheme	Upstr	N hom.	N inhom.	2D	S	S	S
Limiter	–	–	–	–	ψ_{VL}	ψ_1	ψ_2
2(5)	.165	.309	.0867	.107	.0776	.0782	.0750
3(2)	.0844	.148	.0267	.0277	.0215	.0209	.0242
4(2)	.0414	.0717	.00957	.00765	.00596	.00562	.00795
5(0)	.0393	.0683	.00914	.00733	.00572	.00542	.00753
5(1)	.0200	.0332	.00369	.00154	.00136	.00128	.00303
5(2)	.0205	.0347	.00394	.00191	.00144	.00139	.00321
⋮							
5(6)	.0205	.0347	.00393	.00178	.00140	.00136	.00231
	1	2	3	4	5	6	7

Table 6.3: Linear inhomogeneous problem.

Once again the algorithm employing the upstream scheme demonstrates a first order convergence, as resulting from Column 1. Columns 2 and 3 present the experiments with two versions of the N scheme - without weighting of the right-hand-side (we shall call it "homogeneous") and inhomogeneous (which was constructed as an intermediate stage towards the second order accurate S scheme, see Sec.4.3). The homogeneous N

scheme demonstrates first order convergence and its solution error is even larger than that of the upstream scheme. However, the inhomogeneous N scheme leads to the much smaller error and even seem to demonstrate the higher order convergence on coarse grids. This is because the streamwise component is second order small in this case (due to the weighting of the right-hand-side), but it can dominate the cross-stream first order error on some coarse grids. The algorithms based upon the 2D and S schemes (modified for the inhomogeneous case) using any one of the three limiters produce second order accurate solutions. In the cases of the 2D and S schemes with Van Leer's and Roe's ψ_1 limiters this is already clearly achieved by $1FMG$ algorithm. When Roe's ψ_2 limiter is employed the convergence may be just a little slower.

6.1.2 Resolution of contact discontinuities

We shall examine the performances of these algorithms in case of contact discontinuity. Consider Eq.(6.4) with boundary conditions given by

$$u = H(0.5y - x + 1), \tag{6.10}$$

where H is the Heaviside function: $H(x) = 0$ for $x < 0$, $H(x) = 1$ for $x \geq 0$.

It is easy to see that (6.10) is also the exact solution of (6.4) under these boundary conditions, and it contains a jump discontinuity along the line $y = 2x - 2$.

Figures 6.1$(a-f)$ present the numerical solutions to this problem obtained by $2FMG$ algorithms employing different discretization schemes. The results of the $1FMG$ algorithm can hardly be distinguished from those of $2FMG$ algorithm (except the case of the S scheme with Roe's ψ_2 limiter), therefore we omit them. Figures 6.2$(a-f)$ correspond to the same numerical experiments, but present a plot of the solution along the gridline $y = 1.75$ for $1 \leq x \leq 3$. The line with "cross" points represents solutions obtained on level 4, and the line with "diamond" points represents the level 5 solutions - the same as in Figure 6.1.

Figures 6.1*a*, 6.2*a* and 6.1*b*, 6.2*b* correspond to the upstream and N schemes respectively. The contact discontinuity is resolved better in case of the N scheme. However, this result is still unsatisfactory because the width of the transition layer decreases only by factor $\sqrt{2}$ when the grid becomes twice finer, i.e., the number of gridpoints in the transition layer increases by roughly a factor of $\sqrt{2}$, as can be observed in Figure 6.2*b*. Figures 6.1*c* and 6.2*c* correspond to the 2D scheme. The spurious oscillations can be observed in the solution.

Figures 6.1$(d-f)$ and 6.2$(d-f)$ correspond to the S scheme using Van Leer's and Roe's ψ_1 and ψ_2 limiters respectively. The transition layer is slightly wider in case of Roe's ψ_1 than Van Leer's limiter. The discontinuity profile produced by the S scheme with Roe's ψ_2 limiter did not converge enough. This is because the large amount of artificial compression is introduced by such a limiter in the neighborhood of the discontinuity, which badly affects the smoothing properties of the scheme. The solution can be improved by local relaxation sweeps in the neighborhood of the discontinuity.

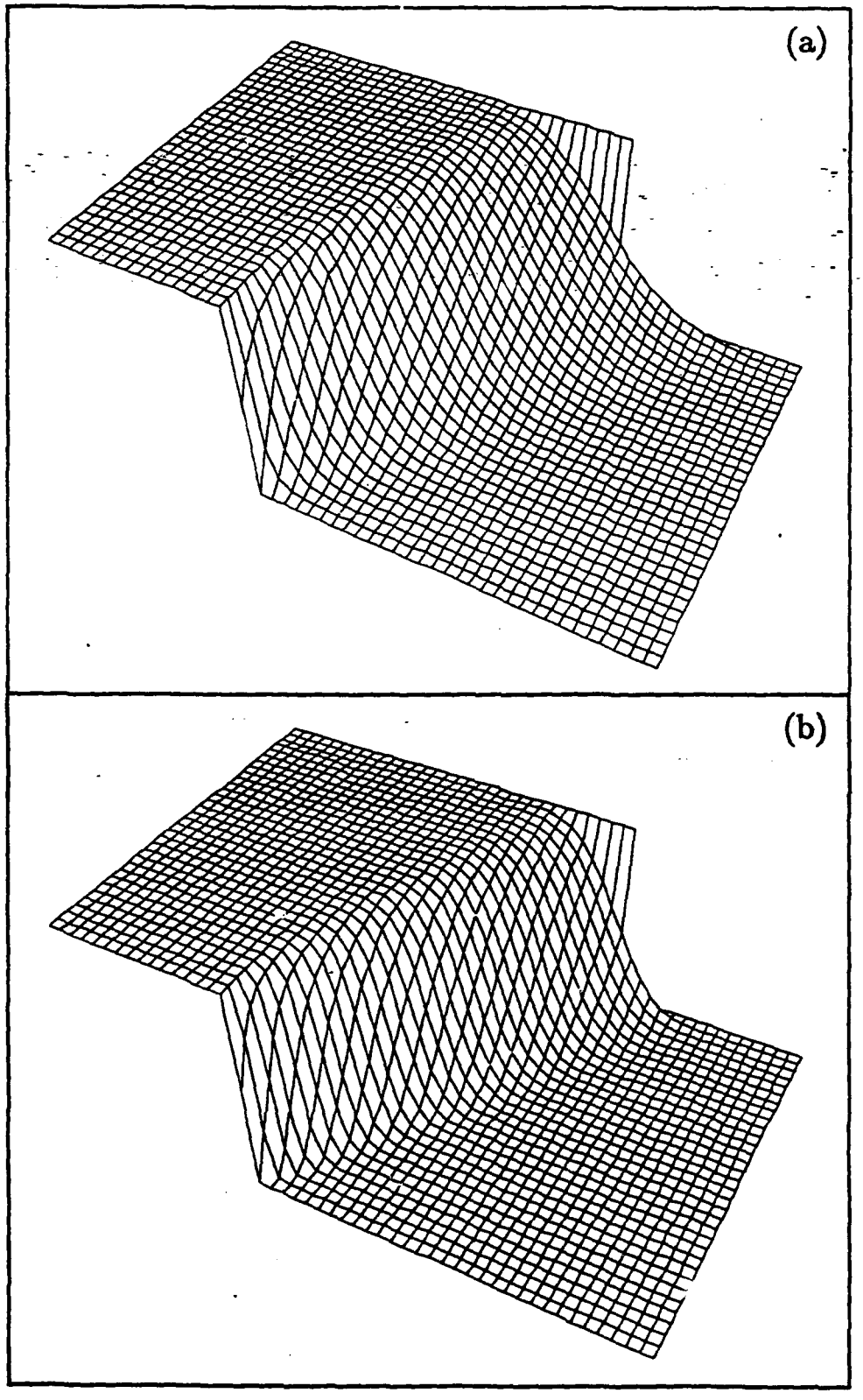

Figure 6.1: Contact discontinuity; (a) - Upstream scheme, (b) - N scheme.

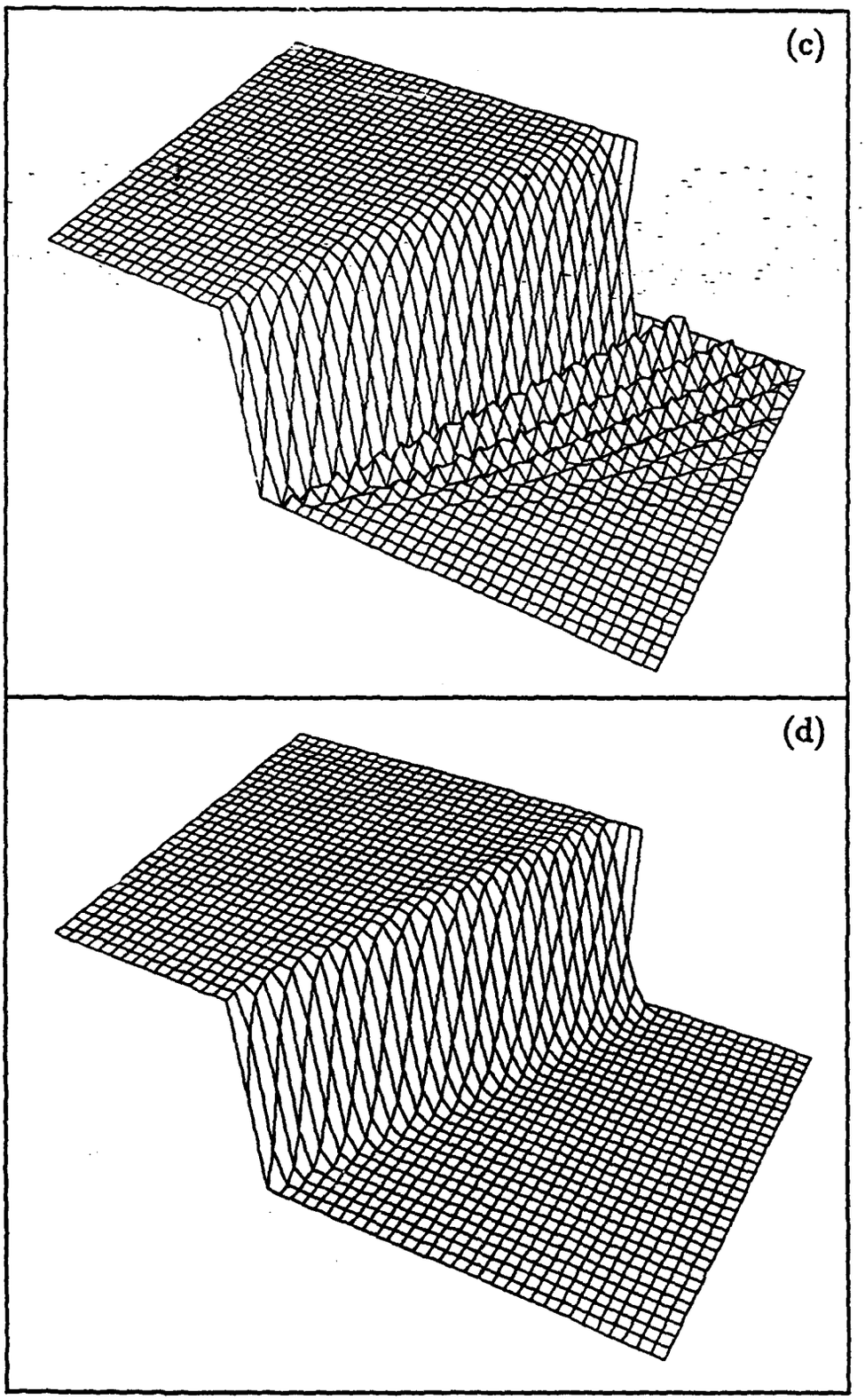

Figure 6.1: continued; (c) - 2D scheme, (d) - S scheme with Van Leer's limiter.

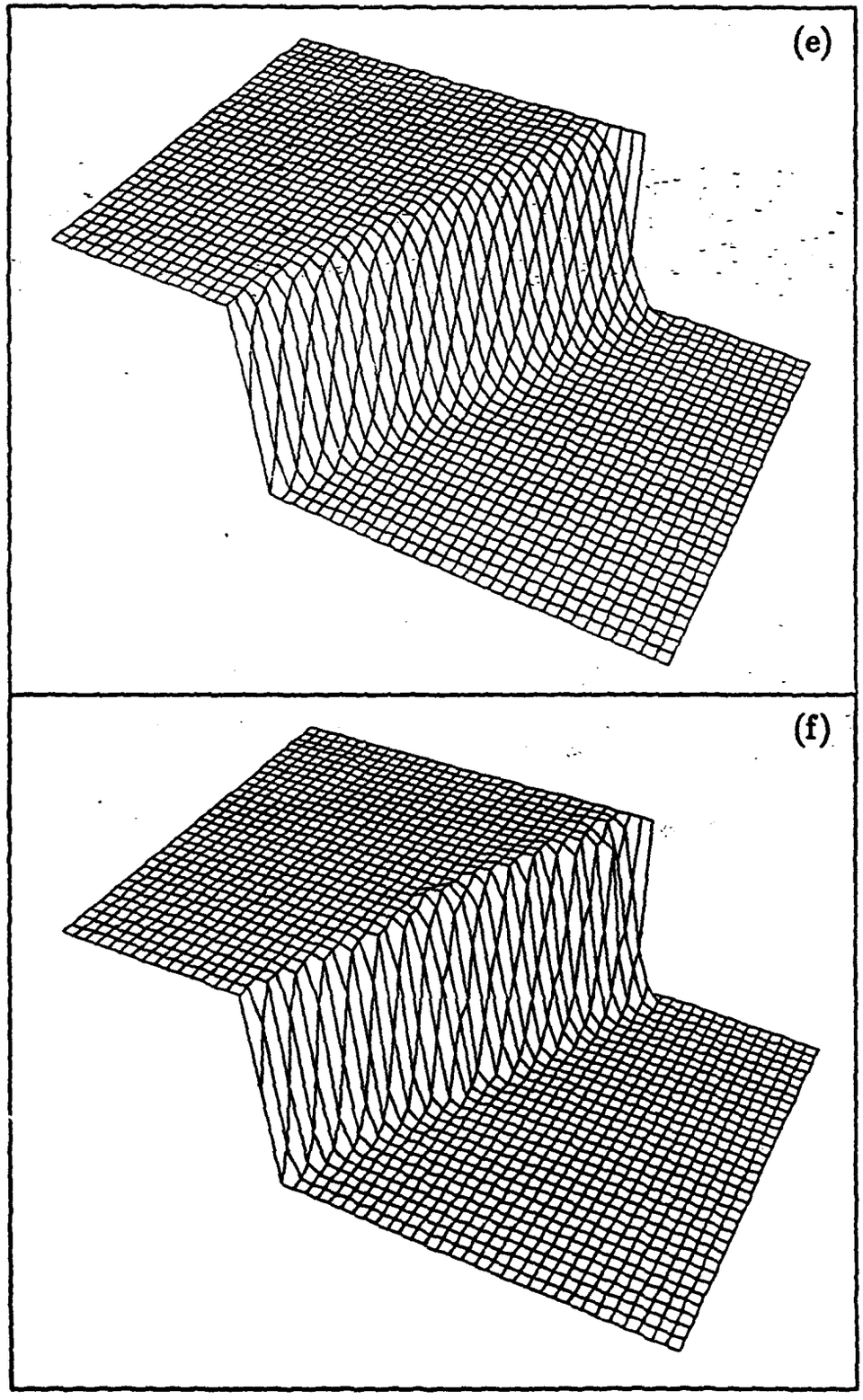

Figure 6.1: continued; S scheme with: (e) - Roe's ψ_1 limiter, (f) - Roe's ψ_2 limiter.

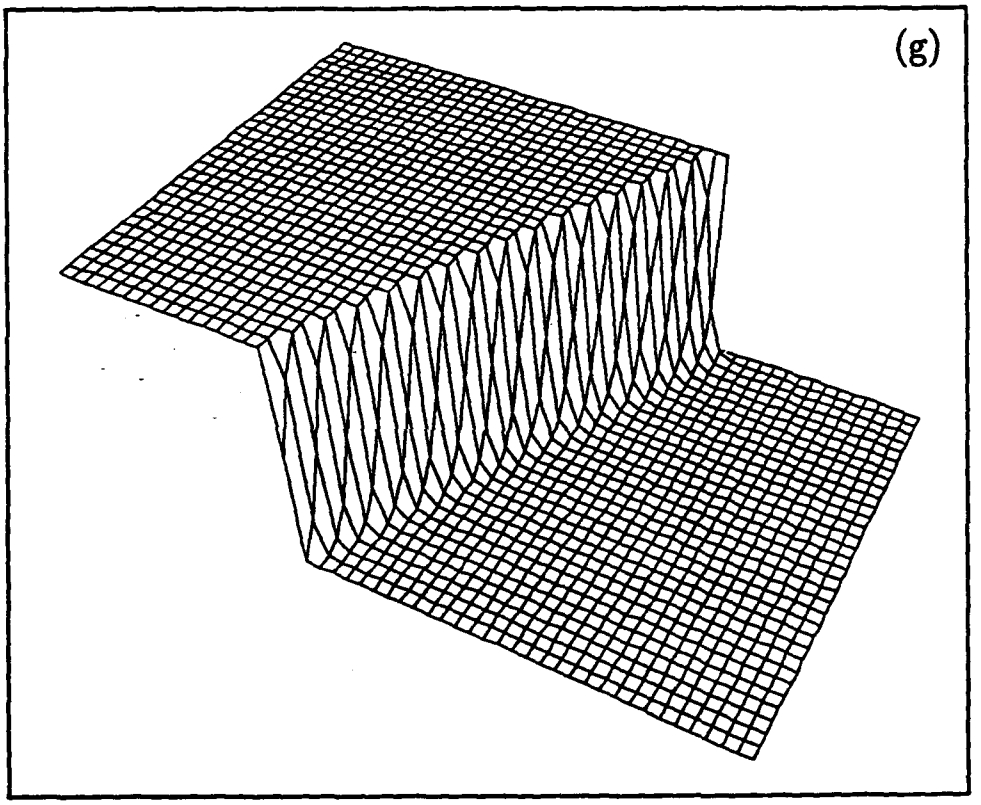

Figure 6.1: continued; (g) - S scheme with Roe's ψ_2 limiter, limit solution.

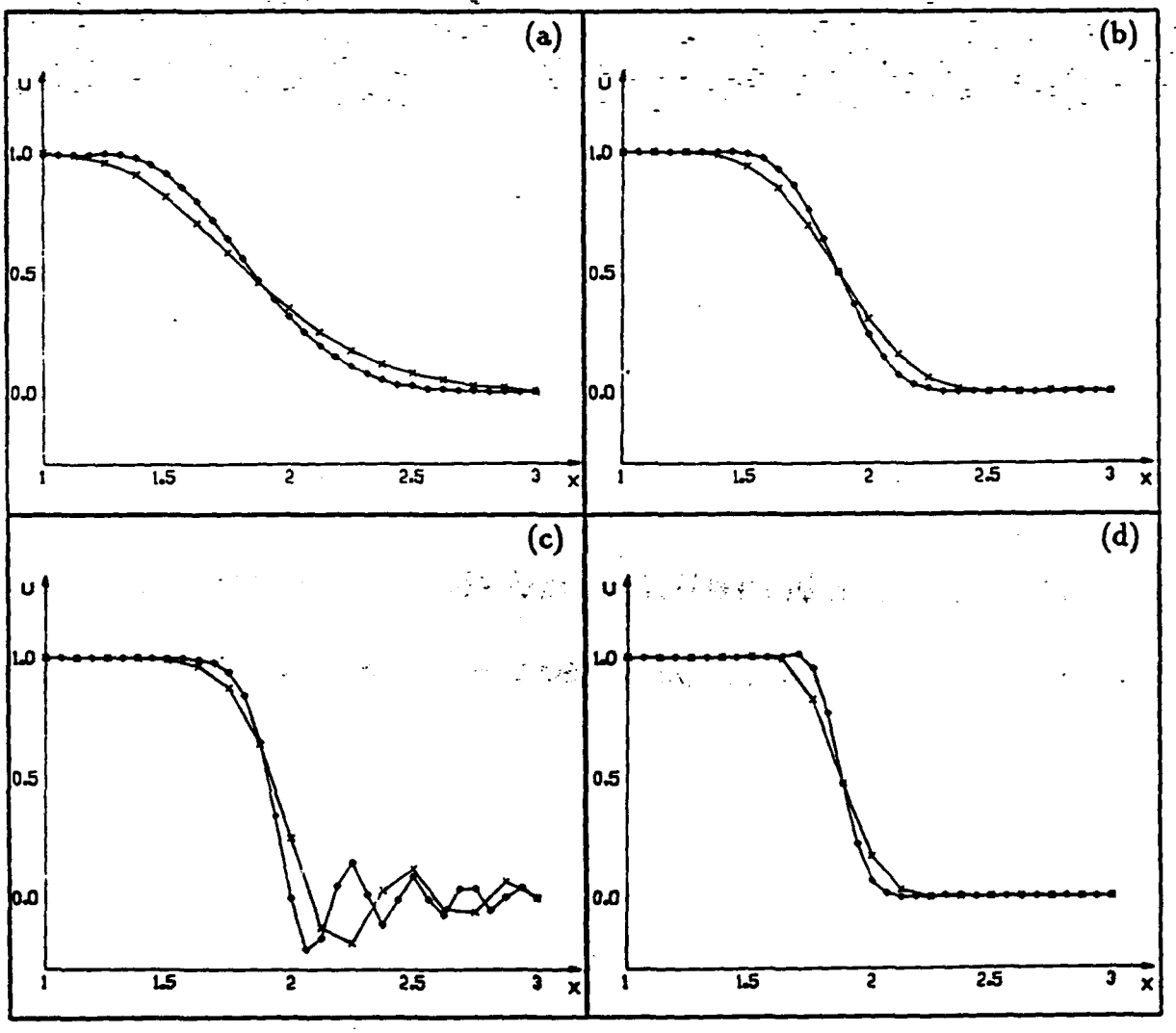

Figure 6.2: Contact discontinuity; gridline $y = 1.25$ for $1 \leq x \leq 3$, $\times$ – level 4, $\diamond$ – level 5: (a) - Upstream scheme, (b) - N scheme, (c) - 2D scheme, (d) - S scheme with Van Leer's limiter.

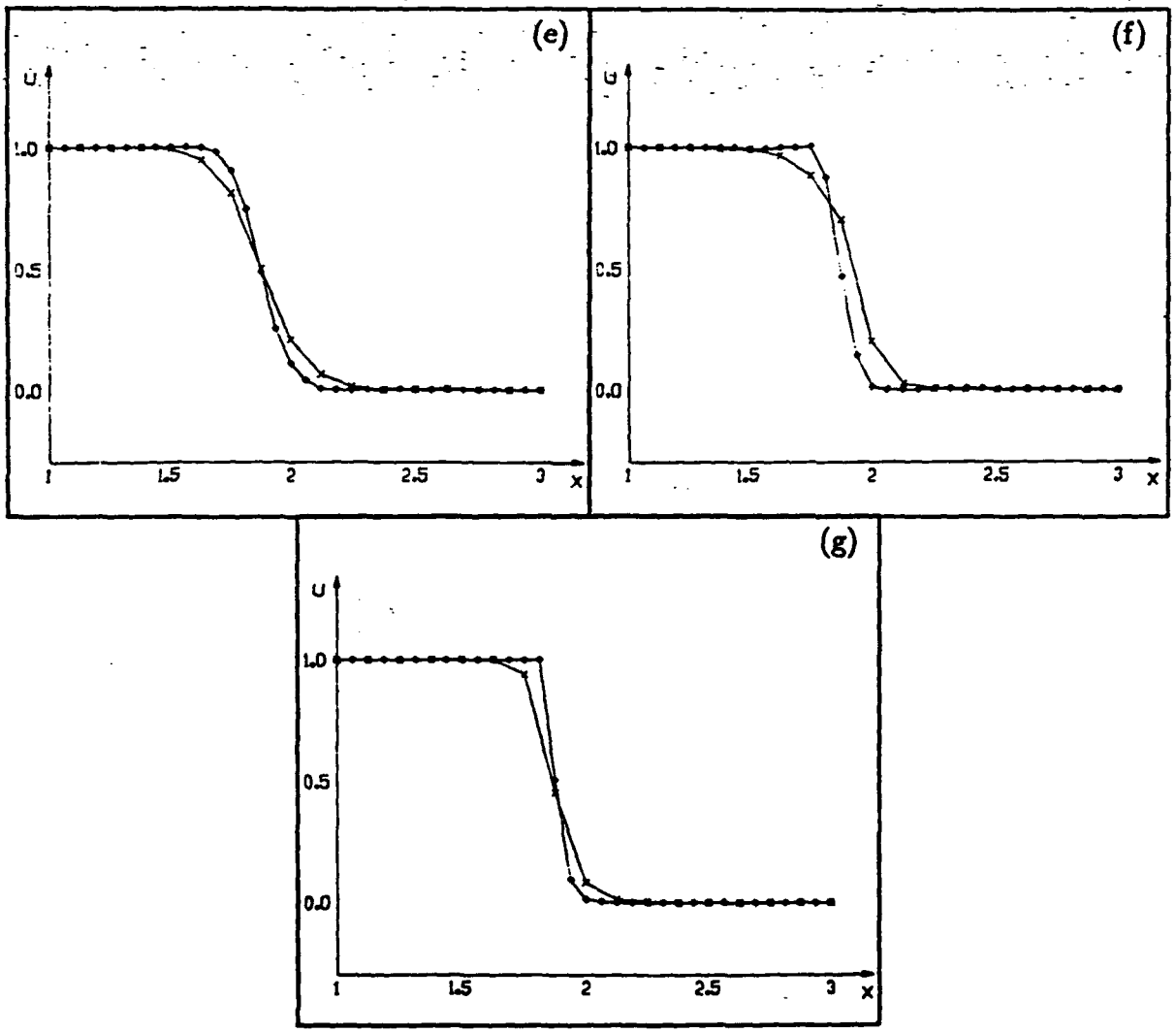

Figure 6.2: continued; S scheme with: (e) - Roe's ψ_1 limiter, (f) - Roe's ψ_2 limiter, (g) - Roe's ψ_2 limiter, limit solution.

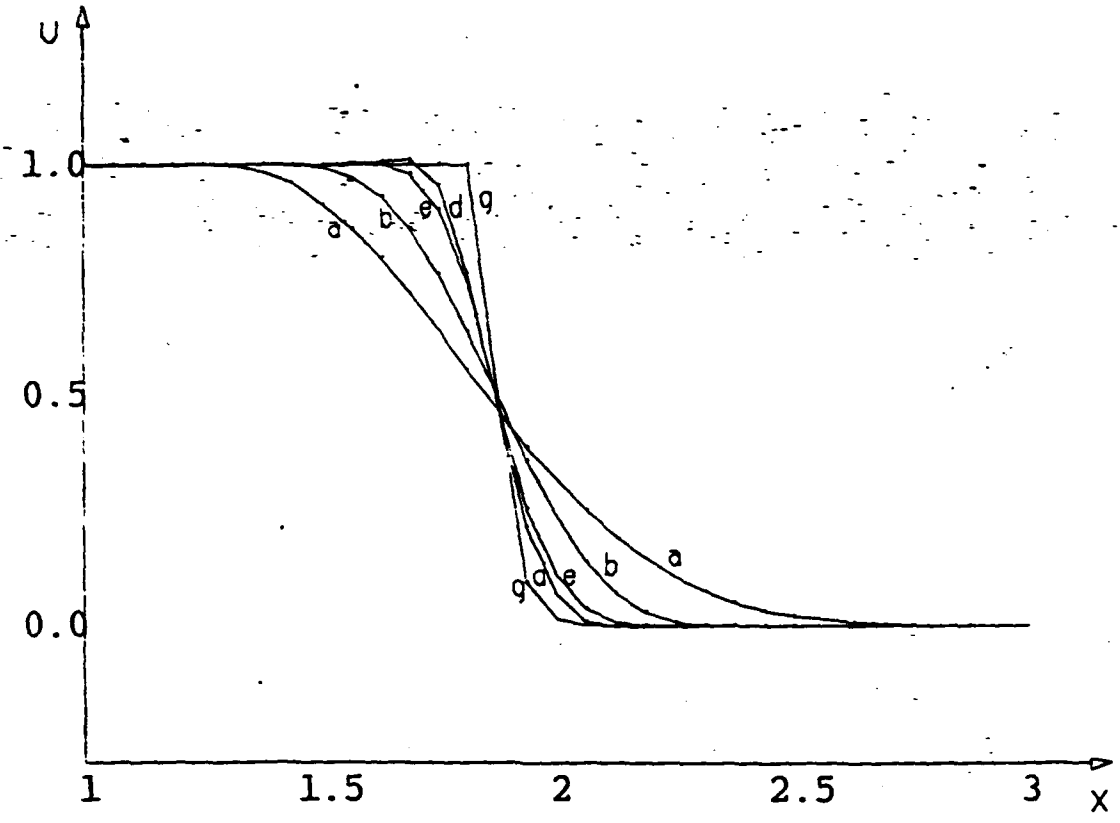

Figure 6.3: Contact discontinuity; (a) - Upstream scheme, (b) - N scheme; S scheme with: (d) - Van Leer's limiter, (e) - Roe's ψ_1 limiter, (g) - Roe's ψ_2 limiter, the limit solution.

Figures 6.1*g* and 6.2*g* present the numerical solution to the same problem obtained by few relaxation sweeps in downstream direction on the finest grid using the S scheme with Roe's ψ_2 limiter. The transition layer created is very sharp. It is possible to obtain the same result by one downstream relaxation sweep, just performing several Newton iterations at each grid point. Note that downstream relaxation sweeps can be performed locally in the neighborhood of the discontinuity together with direction free relaxation all over the domain.

Figure 6.3 presents together the same (level 5) numerical solutions which appear on Figures 6.2*a*, *b*, *d*, *e*, *g*.

The sharpest discontinuity profile here corresponds to the limit solution of the S scheme with Roe's ψ_2 limiter. The S scheme with Van Leer limiter resolves the discontinuity slightly better then with Roe's ψ_1 limiter, but the discontinuity profiles in both cases are sharp (having $O(h)$ width). The first order schemes are obviously inferior to the second order schemes in terms of the discontinuity resolution.

As follows from the argument in Chap.2, the two factors which determine the error in discontinuity location are: the smooth region error at the points L and R (see Figure 2.3) and the error of extrapolation. Since the solution in this example is a piecewise constant, points L and R can be chosen far enough from the discontinuity so that the numerical fluxes at these points will be exact. The extrapolation by constant is exact in this case as well. Therefore, there should be a zero error in the discontinuity location. Indeed, we have observed, that the discontinuity location in the limit solutions in these cases can be recovered upto the round-off error.

6.2 Nonlinear equation

We consider here the following version of Eq.(6.1):

$$-(0+)\Delta u + (u^2/2)_x + u_y = s. \tag{6.11}$$

This equation can be rewritten in the quasilinear form

$$-(0+)\Delta u + uu_x + u_y = s. \tag{6.12}$$

It is easy to see form (6.12) that the characteristic directions are given in this case by the vector $(u, 1)$, i.e. it depends on the solution (For the description of the possible solutions see in Sec.2.1.2).

6.2.1 Smooth solution

In case the boundary conditions are given by (6.8) and

$$s = \cos(x+y)(1+\sin(x+y)) \tag{6.13}$$

it is easy to see that (6.8) is also the exact solution of (6.11).

Although we choose boundary conditions so that there will be no boundary layers created, numerical boundary layers may appear. This is because some of the difference schemes we shall use are not exactly upstream. In order to avoid any influence of potential numerical boundary layers in our study of interior behavior, the error was measured only in the subdomain

$$\Omega' = \{(x,y) : 1/2 \le x \le 5/2, 0 \le y \le 4/3\} \tag{6.14}$$

Numerical experiments with this model problem are presented in Table 6.4.

Again the algorithms employing upstream and homogeneous N schemes produce first order accurate solutions, as can be seen from Columns 1,2. The inhomogeneous N scheme seems to demonstrate higher order convergence on coarser grids because of the same reason as in the case of linear inhomogeneous problem. The $1FMG$ based upon the S1 and S2 scheme leads to second order accurate solutions (see Columns 4-7). The

Difference scheme	Upstr	N hom.	N inhom.	S1	S1	S2	S2	S1&2
Limiter	–	–	–	ψ_{VL}	ψ_1	ψ_{VL}	ψ_2	ψ_1
2(5)	.189	.299	.159	.158	.158	.286	.285	.186
3(2)	.0968	.146	.0438	.0410	.0414	.118	.116	.0524
4(2)	.0477	.0705	.0128	.0105	.0107	.0326	.0318	.0139
5(0)	.0487	.0669	.0134	.0107	.0110	.0334	.0326	.0134
5(1)	.0232	.0331	.00436	.00229	.00227	.00829	.00813	.00359
5(2)	.0241	.0338	.00474	.00271	.00281	.00911	.00889	.00355
⋮								
5(6)	.0242	.0339	.00472	.00265	.00273	.00909	.00889	.00345
	1	2	3	4	5	6	7	8

Table 6.4: Nonlinear inhomogeneous problem; slowly varying solution.

reason for the larger errors demonstrated by the S2 scheme is the additional viscosity (the ϵ correction) which becomes non-zero when the characteristic directions are non-constant. This additional viscosity causes that the S2 scheme demonstrates errors on the coarse grids even larger than those obtained by the first order accurate inhomogeneous N scheme. Note that the choice of the limiter in this problem does not affect the results. Column 8 presents the results obtained by the "blended" S1&2 scheme which contains the S1 scheme with the weight 9/10 and the S2 scheme with the weight 1/10. This scheme demonstrates the second order convergence as each one of its components separatly, however, its errors are much closer to those of the S1 scheme than of the S2 scheme. Note that the S1&2 scheme when used with Roe's ψ_1 limiter is monotonic in case of a homogeneous problem.

Table 6.5 presents numerical experiments with Eq.(6.11) and with the right-hand-side

$$s = 3\cos(3(x+y))(1+\sin(3(x+y))) \tag{6.15}$$

and with the boundary conditions

$$u = \sin(3(x+y)), \tag{6.16}$$

which is the exact solution of this problem as well.

Again the upstream and homogeneous N scheme demonstrate the first order convergence for this problem (Columns 1,2) The inhomogeneous N scheme leads to much smaller errors and seems to be higher order accurate on the coarsers grids (Column 3). The S1 scheme starts to demonstrate second order convergence on the finer levels, but more than 2 cycles on the finest level achieve a second order accurate solution (Columns 4,5) (see also Columns 4-6 in Table 6.2 and Sec.2.3 in [4]). The S2 scheme does not clearly demonstrate the second order convergence even on the finest levels used in this

Difference scheme	Upstr	N (hom.)	N (inhom.)	S1	S1	S2	S2	S1&2
Limiter	–	–	–	ψ_{VL}	ψ_1	ψ_{VL}	ψ_2	ψ_1
2(5)	.668	.780	.615	.611	.613	.573	.573	.553
3(2)	.356	.464	.297	.292	.294	.457	.456	.332
4(2)	.209	.243	.107	.0963	.0971	.303	.300	.125
5(0)	.201	.234	.104	.0935	.0943	.290	.288	.121
5(1)	.125	.121	.0442	.0357	.0363	.180	.177	.0469
5(2)	.121	.119	.0392	.0285	.0295	.124	.121	.0388
⋮								
5(6)	.121	.116	.0360	.0238	.0252	.107	.106	.0341
	1	2	3	4	5	6	7	8

Table 6.5: *Nonlinear inhomogeneous problem; oscillatory solution.*

experiment (Columns 6,7) and leads to much larger errors than the S1 scheme. The S1&2 scheme starts to be second order convergent on the finest levels and leads to errors slightly larger than those of the S1 scheme. The choice of the limiter does not affect the results.

6.2.2 Discontinuous solution

Consider again Eq.(6.11) with $s = 0$.

Shock wave

If we choose (6.10) to be the boundary conditions for this problem, it will be the exact solution for it as well. Figures 6.4$(a-c)$ present plots of the numerical solutions to this problem by the $2FMG$ algorithm employing the S1,S2 and S1&2 schemes respectively.

The choice of limiter in this problem has no influence on the results. The shock profiles in case of the S1 and S1&2 schemes are similar, however the one created by the S2 scheme is less sharp because of the larger cross-stream viscosity due to the e-correction. All the profiles have $O(h)$ width.

Rarefaction wave

Consider the same homogeneous equation as before with boundary conditions

$$u = 1 - 2H(x - 1.5), \tag{6.17}$$

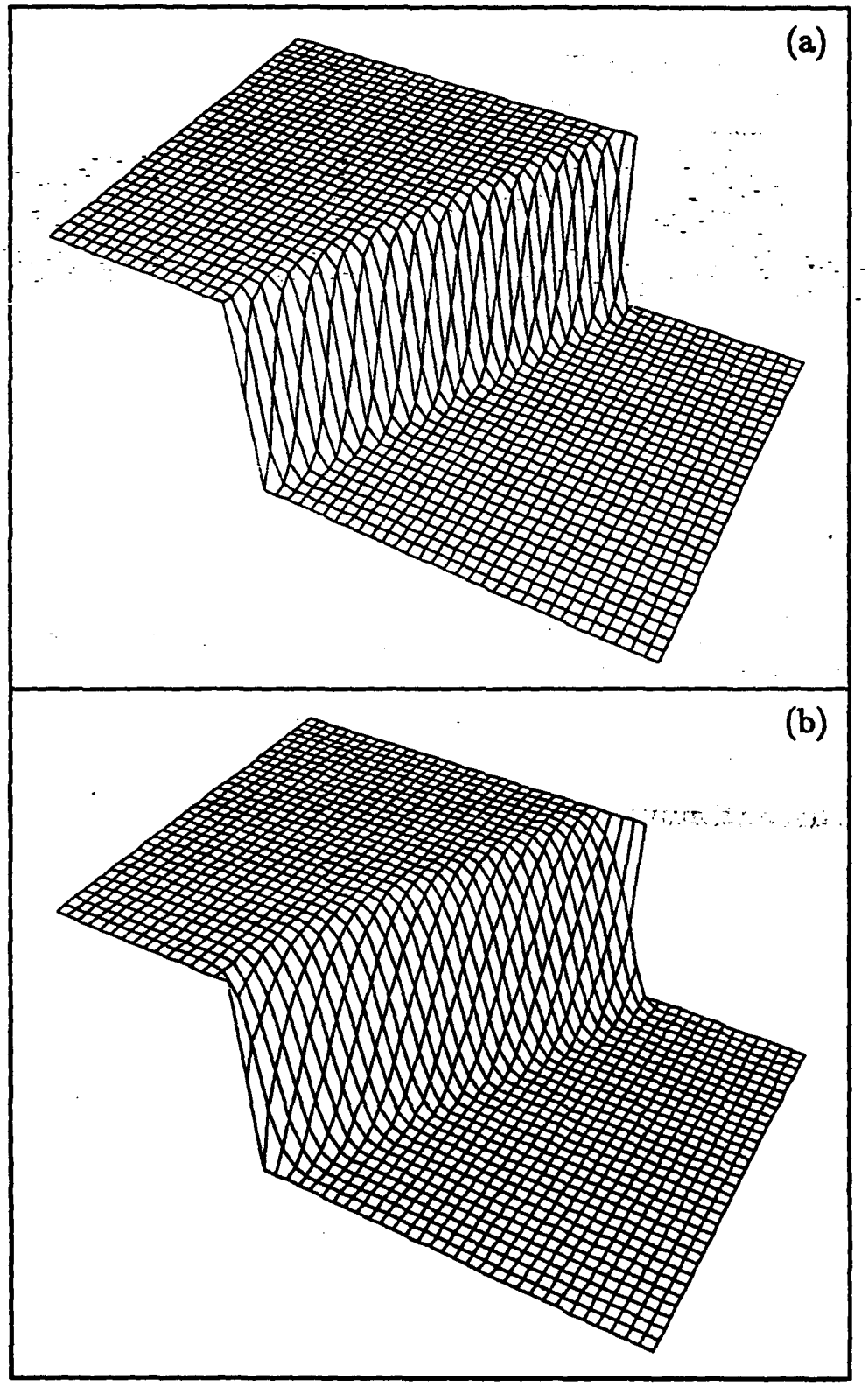

Figure 6.4: Shock wave; (a) - S1 scheme, (b) - S2 scheme.

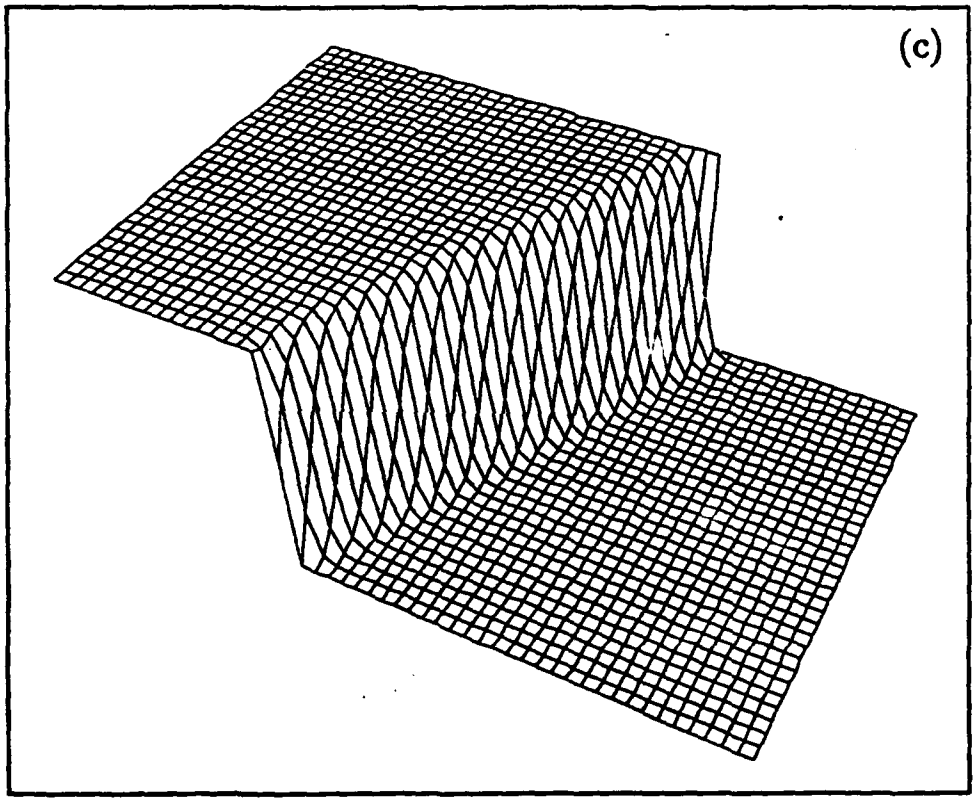

Figure 6.4: continued; (c) - S1&2 scheme.

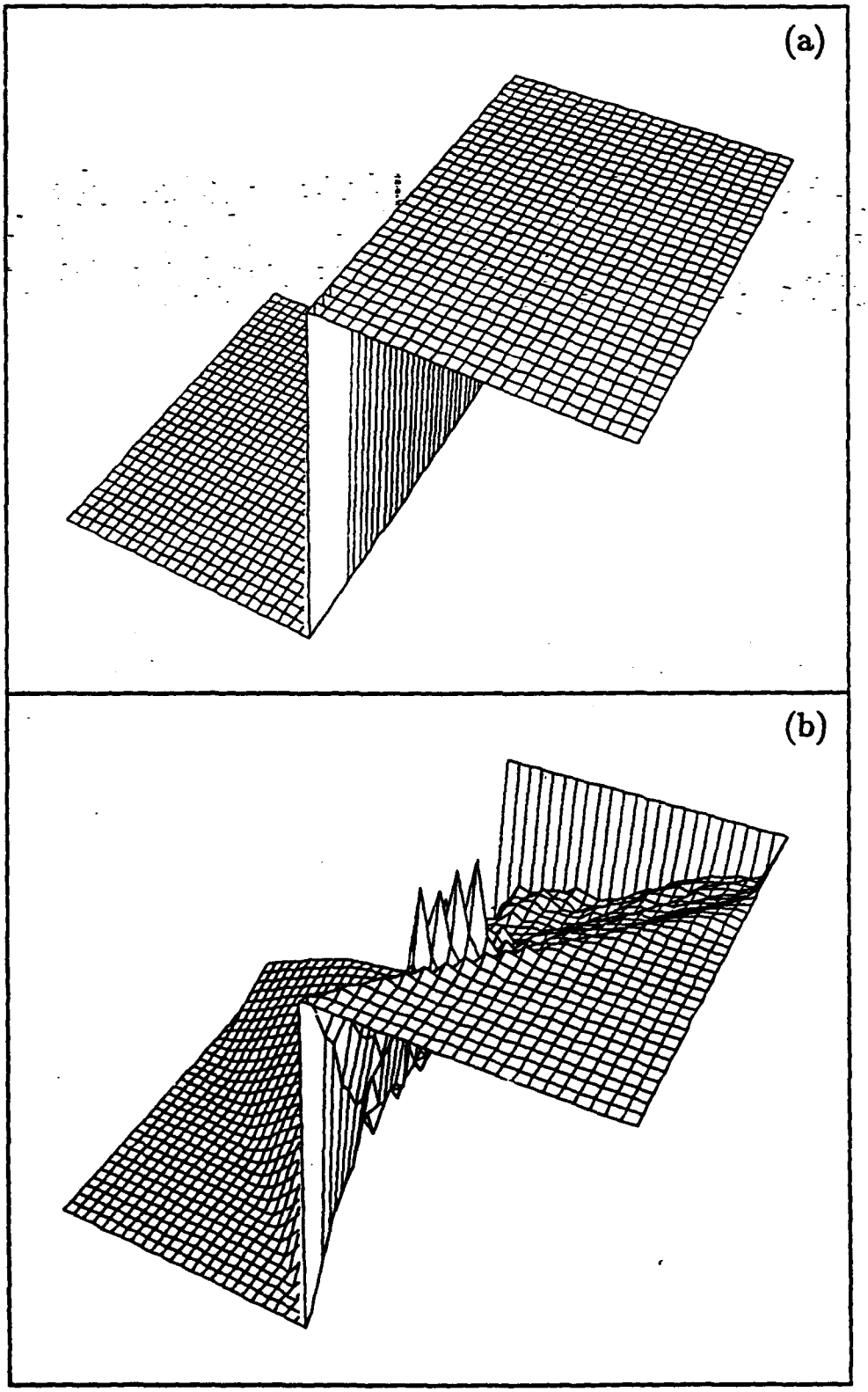

Figure 6.5: Rarefaction wave; (a) - non-physical solution, (b) - S1 scheme with Roe's ψ_1 limiter.

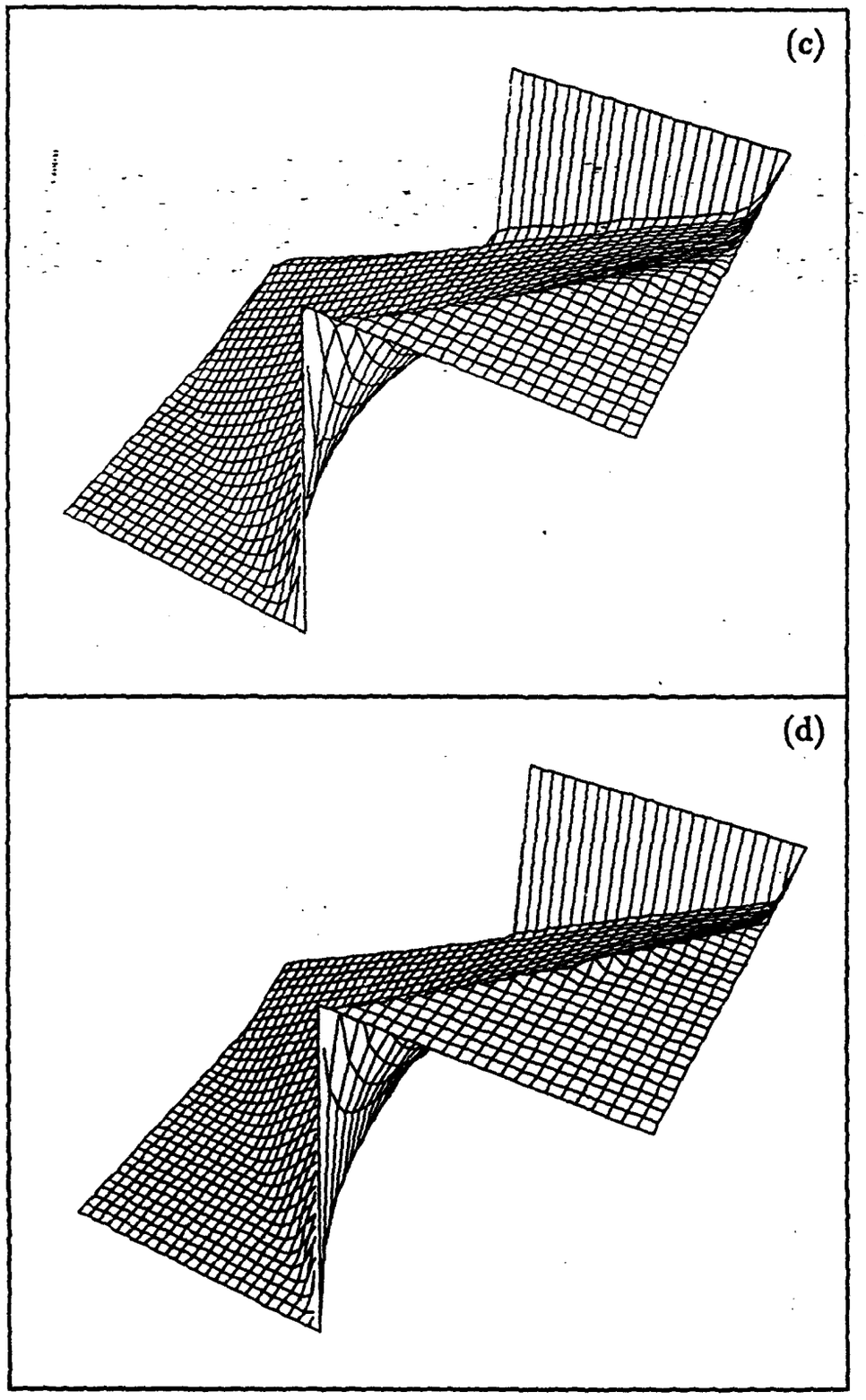

Figure 6.5: continued; (c) - S2 scheme, (d) - S1&2 scheme.

where H is again the Heaviside function. Its plot is presented on Figure 6.5*a*. Such a solution can be obtained also by the upstream, N or S1 schemes (as well as by the Spekrejse scheme). This is because these schemes demonstrate zero residuals on such a grid function. However, this solution is non-physical, it violates the entropy law – there are characteristics which go out from the discontinuity and not from the boundary. Figure 6.5*b* presents the solution to this problem obtained by $2FMG$ algorithm based upon the S1 scheme with ψ_1 limiter. We can observe that an unadmissible discontinuity starts to develop. Figure 6.5*c* presents the solution to this problem obtained by the same algorithm based upon the S2 scheme. (The choice of the limiter does not affect the result in this case as well.) This solution does not contain a non-physical discontinuity. This is due to the e correction which adds some additional viscosity to the scheme when characteristic directions are not constant.

Figure 6.5*d* presents the solution obtained by the $2FMG$ algorithm employing the blended S1&2 scheme. No discontinuity is created as well as in the case of the S2 scheme.

6.2.3 Non-constant solution containing discontinuities

In order to examine the accuracy of the method in smooth regions in the presence of discontinuities and to test the discontinuity locating technique, we shall construct a model problem with a known non-constant solution which contains a shock at a known location. Regarding y as a time-like direction, consider the following hyperbolic problem

$$(u^2/2)_x + u_y = 0, \tag{6.18}$$

with initial conditions given along the x axis by

$$u|_{y=0} = u_0 = .5sin(\pi x) + .5. \tag{6.19}$$

Let us compute the exact solution to this problem. It is constant along characteristics and therefore characteristics are straight lines. Therefore, for every point (x, y) we can find the point $(x_0, 0)$ where the characteristic line which goes through it crosses the x-axis by solving the implicit equation

$$x_0 = x - u_0(x_0)y. \tag{6.20}$$

Of course, the solution is not unique. But if the entropy condition is imposed, or if the non-viscous limit of the viscous equation (6.11) is considered, the solution becomes unique. It is easy to see that inside the rectangle (6.2) it will contain one shock wave (see Figure 6.6) going along the line

$$y = 2x - 2. \tag{6.21}$$

Consider again the domain (6.2), and let (6.19) be the boundary conditions for Eq.(6.11).

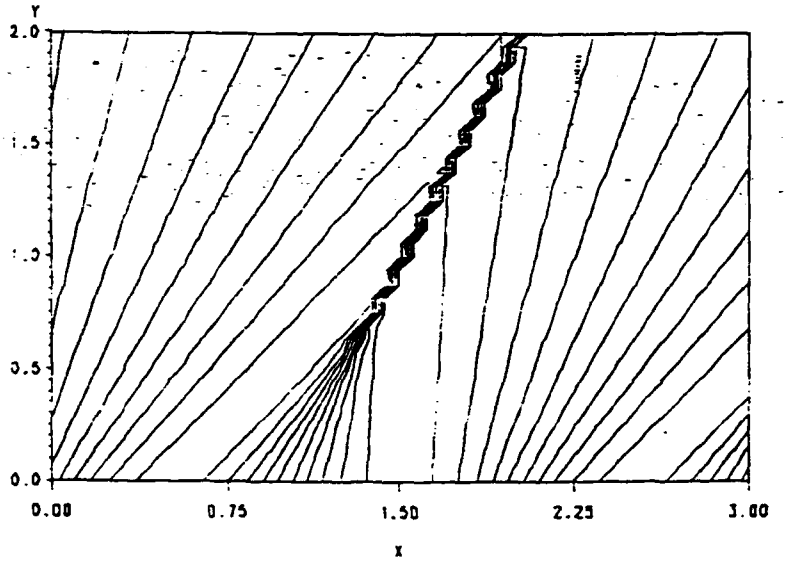

Figure 6.6: Characteristic lines and shock wave in the exact solution.

Differ. scheme	S1		S1		S2		S1&2	
Limiter	ψ_1		ψ_{VL}		ψ_{VL}		ψ_{VL}	
Error	Sol.	Sh.Loc	Sol.	Sh.Loc	Sol.	Sh.Loc	Sol.	Sh.Loc
2(5)	.0621	–	.0556	–	.150	–	.0632	
3(2)	.0100	.105	.00765	.111	.0622	.200	.0136	.0969
4(2)	.00304	.0141	.00274	.0165	.0188	.0240	.00405	.0109
5(0)	.00297	.00427	.00266	.00362	.0188	.0134	.00383	.00654
5(1)	.00089	.00283	.00090	.00246	.00642	.0116	.00112	.00131
5(2)	.00085	.00215	.00069	.00224	.00603	.00827	.00114	.00184
⋮								
5(6)	.00078	.00225	.00069	.00243	.00594	.00591	.00108	.00159
	1		2		3		4	

Table 6.6: Nonlinear problem; varying solution containing a shock wave.

The numerical experiments presented by Table 6.6 and Figure 6.7 are performed with this model problem. The $2FMG$ algorithm is used. In order to avoid any influence of numerical boundary layers and the shock we measure the solution error L_1 norm in the following subdomain

$$\Omega'' = \{(x,y) : 1/2 \le x \le 5/2, 0 \le y \le 2x-3 \; or \; \max(0,2x-1) \le y \le 4/3\}. \quad (6.22)$$

The intersection point between the shock and a grid line parallel to the x axis is calculated according to the method described in Chap.2. The segment LR is taken to be 6 meshsizes long with its middle close to the assumed shock location. The extrapolation by a constant is used. The error in shock location is measured for $1 \le y \le 2$, where the discontinuity is strong enough.

Each pair of columns in Table 6.6 presents one experiment. The first column in each pair displays the history of the L_1 norm of the solution error and the second - the L_1 norm of the shock location error. Again we can conclude that a second order accurate solution (in smooth regions and also in terms of the discontinuity location) can be obtained by the $2FMG$ algorithm using either the S1 or the S2 schemes. However, the S1 scheme leads to smaller errors than the S2 scheme. The choice of the limiter does not affect the solution. The S1&2 scheme leads to a solution error slightly larger than the S1 scheme, but to a smaller error in the discontinuity location. Figures 6.7($a-c$) present plots of the numerical solution obtained by the $2FMG$ algorithm in the same cases which are presented in Columns 2-4 of Table 6.6.

The solution plots corresponding to the S1 scheme with Roe's ψ_1 (Column 1) and Van Leer's (Column 2) limiters are undistinguishable. Therefore, we omit the first one. In spite of the fact that the monotonicity of the solution is not guaranteed when the S1 scheme is used with Van Leer's limiter, we do not observe any oscillations on Figure 6.7a. Also we can see that the S2 does not provide the shock resolution as sharp as the S1 scheme. The $2FMG$ algorithm which uses the S1&2 scheme leads to the solution presented by Figure 6.7c. The shock resolution is comparable with that of the S1 scheme.

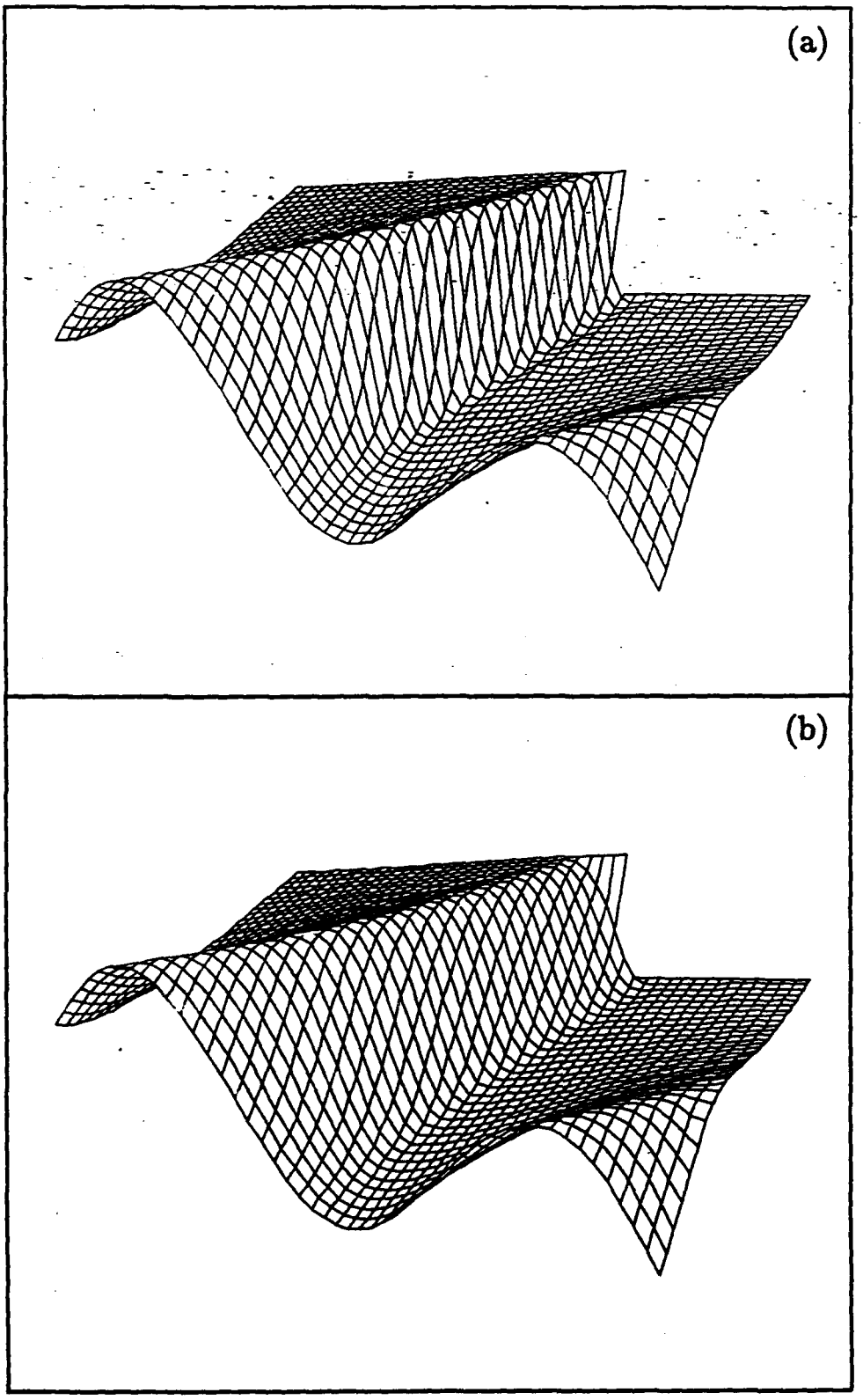

Figure 6.7: Non-constant solution containing a shock wave; (a) - S1 scheme with Roe's ψ_1 limiter, (b) - S2 scheme with Roe's ψ_2 limiter.

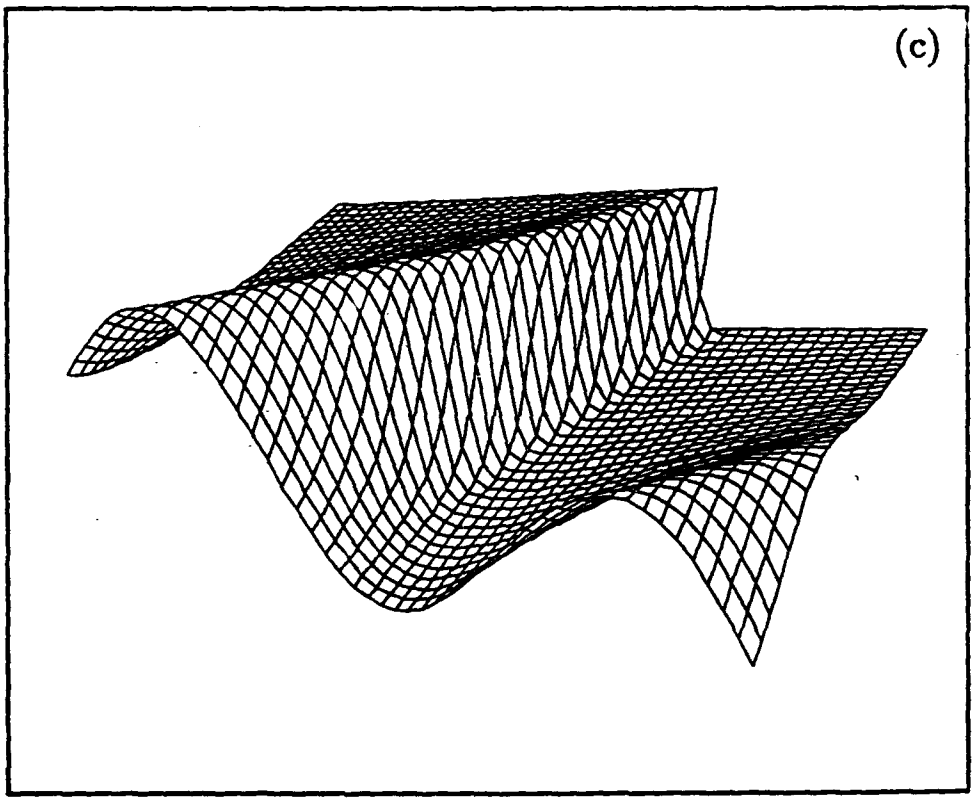

Figure 6.7: continued; (c) - S1&2 scheme with Roe's ψ_1 limiter.

6.3 Choice of discretization

Both S1 and S2 (as well as the blended S1&2 scheme) are second order accurate schemes when used with any of the limiters considered above.

In case we want to obtain smaller errors in smooth regions, sharper shock resolution and better discontinuity location – the S1 scheme should be used. The S1 scheme is also proven to produce monotonic solution when used with Roe's ψ_1 limiter in case of homogeneous problems. However, non-physical discontinuities can appear in the solution in case of a rarefaction wave aligned with the grid.

The S2 scheme is proven to be monotonic with any of the limiters considered above. It does not admit non-physical discontinuities. However, it leads to much larger solution and discontinuity location errors and provides shock resolution inferior to that of the S1 scheme.

An interesting possibility is therefore to use the blended S1&2scheme, i.e. to average the S1 and S2 schemes with certain weights. This allows us to achieve a more accurate solution and better resolution of discontinuities but still without admitting the non-physical ones. The S1&2 scheme is monotonic in case of homogeneous problems when used with Roe's ψ_1 limiter. Another alternative can be to use the S2 scheme for the first multigrid cycle on each level and the S1 scheme for the second one. The solution error and discontinuity resolution in this case will be very similar to those of the S1 scheme. However, non-physical discontinuities will not develop.

The choice of a limiter does not influence significantly the resolution of shocks and the solution error for both S1 and S2 schemes. In case better resolution of contact discontinuities is desirable, the compressive limiters (Van Leer's or Roe's ψ_2) should be used. However, sharp edges in the solution can be created because of the artificial compression. When the S1 scheme is used with a compressive limiter, the monotonicity of the solution is not guaranteed.

We can conclude that the choice of discretization may depend on which features of the solution are of more interest to us.

Chapter 7

Discussion and conclusions

7.1 Summary

A fast multigrid solver for a scalar 2D steady-state conservation law is developed.

Two main problems were solved in order to achieve the goal of this work:

1. A new genuinely 2D discretization scheme based on the compact "9-point box" stencil was constructed. This scheme provides a separation between treatment of streamwise and cross-stream directions. The artificial viscosity is added in the streamwise direction only. High resolution can be introduced in the cross-stream direction. As a result, numerical solutions produced by these schemes are monotone with sharp resolution of oblique discontinuities and a local relaxation process applied with this type of discretization is stable. The solutions also demonstrate second order convergence in smooth regions in the case of a homogeneous problem. This property is extended to inhomogeneous equations by a simple weighting of the right-hand side.

2. We have shown that a captured-shock solution (provided the discrete scheme employed is conservative) contains information about the discontinuity location up to the same order of accuracy which is actually achieved in the smooth regions. This assertion has two important implications. The first is the possibility of performing a post-processing: once the discontinuity is recognized in the solution, its location can be extracted with a higher-order accuracy. The second implication is that the usual multigrid correction interpolation (bilinear etc.) can be used in the neighborhood of discontinuities as well as in the smooth regions. Indeed, it expresses the correct movement of the discontinuity provided a conservative residual transfer (Full Weighting) is used.

Numerical experiments confirm that second order accurate (both in smooth regions and discontinuity location) solutions can be usually obtained by the 2FMG algorithm, even with direction-free relaxation. The limit solutions can also be obtained by just few downstream relaxation sweeps on the finest level.

7.2 Remark on double discretization

Our original intention was to use the double discretization technique, i.e. to employ two different schemes in the relaxation and in the residual calculation. The scheme used in

the relaxation process must be stable, but may be low order accurate. The scheme used for the residual calculation must be higher order accurate, but may be unstable. This approach is usually good for smooth solutions. However, there is a certain difficulty to apply it in the presence of discontinuities. Two different schemes lead usually to solutions whith have different discontinuity profiles. Therefore, the residuals calculated by means of one scheme on the solution approximation obtained by relaxing another scheme may attain large values in the neighborhood of discontinuities. If we transfer these large residuals together with others to the coarser grid, the numerical process may become divergent. However, in case both difference schemes used are conservative the summation of the residuals inside the transition layer along gridlines crossing the discontinuity will give small numbers. Therefore, if this "residual cancelling" is performed each time before going to the coarser grid, the multigrid algorithm employing the double discretization will produce higher order accurate solution in smooth regions as well as higher order error in discontinuity location. However, such an algorithm requires to detect transition layers representing discontinuities in the solution and to perform residual cancelling accross these layers on each grid before transfering residuals to the coarser one. This is the reason why we decided to use the same higher order accurate and stable scheme both for the relaxation and the residual calculation.

7.3 Extension to Euler equations

Our next objective is to extend these methods to the steady-state Euler equations. This is not expected to be complicated, because a discretization of the convection operator is the main concern of the approach of [3] for the case of this system too. Moreover, it seems to be possible to design a stable local relaxation process for the Euler system, based on the ideas of [3].

7.4 Efficiency comparison

We shall compare now the number of operations neccessary to perform in order to evaluate a residual at one grid point of the Spekreijse scheme and of the S1 and S2 schemes developed in this work. The results of the comparison are summarized in Table 7.1.

We can conclude, that the S1 and the Spekreijse schemes are comparable in terms of efficiency for residual calculation. The S2 scheme is more expensive. However, we want to note the following:

- The schemes constructed in this work are based on narrow stencils. Therefore, they introduce smaller cross-stream viscosity than the schemes of the same order of accuracy based on the "dimensional splitting" approach (see Sec.2 in [4]). This means that the solution of a certain quality can be obtained by our schemes on a grid coarser than the one needed for the Spekreijse scheme.

Difference scheme	Spekreijse	S1	S2
f & g evaluation	8	5	5
a & b evaluation	8	12	12
r.h.s. evaluation	1	1+4	1+4
limiter evaluation	8	2(4)	2(4)
gridpoint index calc.	9	9+4	9+4
other floating-point operations	25	26(44)	54(72)

Table 7.1: Efficiency comparison: after the "+" sign - the number of additional operations needed to achieve second order accuracy in case of an inhomogeneous problem; in the parenthesis - the number of operations which may be required in the neighborhood of discontinuities; otherwise given is the number of operations required in the smooth regions.

- A pointwise relaxation with the S1 and S2 schemes is stable. To perform a Newton iteration at one point requires (in addition to the residual calculation) the evaluation of the numerical fluxes derivatives with respect to the central point value. This is much less expensive than the residual calculation (especially if the approximate formulas for derivatives are used). The evaluation of the numerical fluxes derivatives in case of the Spekreijse scheme is approximately as expensive as in case of the S1 and S2 schemes. However, a 4-point block relaxation is needed when relaxing the Spekrejse scheme (see [20]). This means, that one Newton iteration should be performed for a system of 4 nonlinear equations each time, instead of being applied for 4 separate equations as in the S1 or S2 schemes. This requires to invert a 4×4 matrix for every 4 gridpoints. This matrix contains at least 4 zero elements, and upto 8 in case of a smooth solution. This additional matrix inversion increases the computational cost of the Spekreijse scheme when comparing to the S1 and S2 schemes.

- The really important point here is that there was not found any (neither local nor global) stable relaxation which can be applied with the Spekreijse scheme in case of the steady-state Euler system. Therefore, it was possible to achieve the second order accuracy only by employing a defect correction method (see [19]), which is not a fully-efficient way to deal with non-elliptic problems (see Sec.2 in [4]). Indeed, many cycles were required for the Spekreijse scheme to reach second order accuracy.

We can conclude that the multigrid solver for a scalar 2D conservation law based on the discretization schemes developed here is slightly more efficient than the one based on the Spekreijse scheme. However, when extended to the Euler system they are expected to be much more efficient than the solver developed in [19].

7.5 Some properties and future development

The versions of the S1 and S2 schemes which were shown to be monotonic for the case of a homogeneous equation seem to be TVD (total variation diminishing). Although we do not prove this, it must follow from the monotonic property for these schemes exactly as in the 1D case. This means that, when these schemes are used to solve a time-dependent problem, convergence of the solution is theoretically justified for a finite time interval. This does not contradict the result of [9], because when used to discretize a time-dependent problem, both the S1 and the S2 schemes will retain only first order spatial accuracy. However, discontinuities will be sharply resolved in the solution. The second order accuracy can possibly be obtained by treating the time derivative as the right-hand-side of an inhomogeneous problem (see Sec.5), but such a discretization will not have the TVD property anymore.

Note that the use of TVD schemes has a theoretical advantage for time-dependent problems and finite time evolution only. However, when constructing a certain scheme

for practical purposes it may be enough to require that it can be rewritten as to another scheme which is not necessarily of the positive type, i.e. it can have small negative coefficients ($O(h)$ compared with the positive coefficients) at some points.

We can also distinguish between smooth regions and the neighborhood of discontinuities, and use different schemes for each of them: a more complicated scheme with limiters near discontinuities only, and a simplified scheme in the smooth regions. These simplifications can lead to a substantial computational cost reduction of the method in smooth regions. Also, additional local relaxation sweeps may be performed in the neighborhood of strong discontinuities, increasing the efficiency of the method for little extra cost.

It must be possible to construct a third-order accurate upstream scheme based on the same "9-point box" stencil: 4 upstream points are enough to obtain an $O(h^3)$ cross-stream truncation error, while the third order accuracy in other directions can be achieved again by a certain weighting of the right-hand-side of the equation. An approach similar to ENO (essentually non-oscillatory) in [13] can be introduced in the cross-stream direction. However, some technical details still have to be verified.

Another interesting possibility is to construct upstream difference schemes based on thiner, but longer stencils. This approach will improve resolution of characteristic components and discontinuities and can possibly lead to a very high order accuracy compact schemes.

Bibliography

[1] Van Albada, G.D., Van Leer, B., Roberts, W.W.,Jr., A comparative study of computational methods in cosmic gas dynamics, ICASE, Report No. 81-24 (1981).

[2] Brandt, A., Multigrid solvers for non-elliptic and singular perturbation steady-state problems, The Weizmann Institute of Science, Rehovot, Israel, 1981.

[3] Brandt, A., Multigrid techniques: 1984 Guide with applications to fluid dynamics, The Weizmann Institute of Science, Rehovot, Israel, 1984.

[4] Brandt, A., The Weizmann Institute research in multilevel computation: 1988 Report, Proc. 4th Copper Mountain Conference on Multigrid Methods, *SIAM*, to appear.

[5] Boerstoel, J.W., Kassies, A., Integrating multigrid relaxation into a robust fast-solver for transonic potential flows around lifting airfoils, AAIA 6th CFD Conference, 1984.

[6] Davis, S., A rotationally biased upwind difference scheme for the Euler equations, ICASE, Report No. 83-87 (1983).

[7] Davis, S., Shock capturing, ICASE, Report No. 85-25 (1985).

[8] Eckhaus, W., Boundary layers in linear elliptic singular perturbation problems, *SIAM Review*, 14 (1972), 225-270.

[9] Goodman, J.B., LeVeque, R.J., On the accuracy of stable schemes for two dimensional conservation laws, *Math. Comp.* 45 (1985), 15-21.

[10] Guckenheimer, J., Shocks and rarefactions in two space dimensions, *Archive for Rational Mechanics and Analysis* 59 (1975), 281-291.

[11] Harten, A., The artificial compression method for computation of shock and contact discontinuities: III. Self-adjusting hybrid schemes, *Math. Comp.* 32 (1978), 363-389.

[12] Harten, A., High resolution schemes for hyperbolic conservation laws, *J. Comp. Phys.* 49 (1983), 357-393.

[13] Harten, A., ENO schemes with subcell resolution, ICASE, Report No. 87-86.

[14] Van Leer, B., Towards the ultimate conservative difference scheme, II. Monotonicity and conservation combined in a second order scheme, *J. Comp. Phys.* 14 (1974), 361-370.

[15] Osher, S., Solomon, F., Upwind difference schemes for hyperbolic systems of conservation laws, *Math. Comp.* **38** (1982), 339-374.

[16] Osher, S., Chakravarthy, S., High resolution schemes and the entropy conditions, ICASE, Report No. 172218, 1983.

[17] Sidilkover, D., Multigrid solvers for singular perturbation steady-state problems, M.Sc. Thesis, The Weizmann Institute of Science, Rehovot, Israel, 1983.

[18] Sidilkover, D., Numerical solution to steady-state problems with discontinuities, Ph.D. Final Report, The Weizmann Institute of Science, Rehovot, Israel, 1988.

[19] Spekreijse, S., Multigrid solution of the steady Euler equations, Ph.D. Thesis, CWI, Amsterdam, Netherlands, 1987.

[20] Spekreijse, S., Multigrid solution of monotone second-order discretization of hyperbolic conservation laws, CWI, Report NM-R8611, 1986.

[21] Sweby, P.K., High resolution schemes using flux limiters for hyperbolic conservation laws, *SIAM J. Numer. Anal.*, **21** (1984), 995-1011.

[22] Vol'pert, A.I., The spaces BV and quasilinear equations, *Math. USSR – Sbornik*, **2** (1967), 225-267.

[23] Woodward, P., Colella, P., The numerical simulation of two-dimensional flow with strong shocks, *J. Comp. Phys.* **54** (1984), 115-173.